NASA Conference Publication 2341

IN
(1+33)

Twentieth Annual Conference on Manual Control

June 12-14, 1984 Ames Research Center, Moffett Field, California

Compiled by: Sandra G. Hart and Earl J. Hartzell

Volume II

(NASA-CP-2341-Vol-2) TWENTIETH ANNUAL CONFERENCE ON MANUAL CONTROL, VOLUME 2 (NASA) 423 p HC A18/MF A01 CSCL 05H
N85-14535
THRU
N85-14568
Unclas
G3/54 20592

NASA
National Aeronautics and
Space Administration

NASA Conference Publication 2341

Twentieth Annual Conference on Manual Control

June 12-14, 1984 Ames Research Center, Moffett Field, California

Compiled by: Sandra G. Hart and Earl J. Hartzell
Ames Research Center
Moffett Field, California

Volume II

NASA
National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

FORWARD

Volumes I and II contain the Proceedings of the Twentieth Annual Conference on Manual Control. The proceedings were published with the support of the National Aeronautics and Space Administration and the Army Aeromechanics Laboratory, both located at Ames Research Center. The meeting was held at the Sheraton-Sunnyvale Hotel in Sunnyvale, California from June twelfth through fourteenth, 1984. Both formal papers that represented completed work and informal papers that represented work in progress were presented. The two volumes include all of the papers accepted for presentation at the meeting; seventy six complete manuscripts and nine abstracts. The papers are divided into two volumes that represent the two general classes of topics that were covered. Volume I covers more traditional "Annual Manual" topics such as time series modeling, flying qualities, and supervisory control models. Volume II contains papers that are more focused on psychological and physiological issues, such as evoked potential and workload measurement, that were included in the program of the concurrent "Annual Mental"

This was the twentieth in a series of conferences dating back to December 1964. These earlier meetings and their proceedings are listed below:

First Annual NASA-University Conference on Manual Control, the University of Michigan, December 1964. (Proceedings not printed.)

Second Annual NASA-University Conference on Manual Control, University of Southern California, February 28 to March 3, 1967. (NASA-SP-128)

Third Annual NASA-University Conference on Manual Control, University of Southern California, March 1-3, 1968. (NASA-SP-144)

Fourth Annual NASA-University Conference on Manual Control, University of Michigan, March 21-23, 1968. (NASA-SP-192)

Fifth Annual NASA-University Conference on Manual Control, Massachusetts Institute of Technology, March 27-29, 1969. (NASA-SP-215)

Sixth Annual Conference on Manual Control, Wright-Patterson AFB, Ohio, April 7-9, 1970. (AFIT/AFFDL Report, no number)

Seventh Annual Conference on Manual Control, Univeristy of Southern California, June 2-4, 1971. (NASA-SP-281)

Eighth Annual Conference on Manual Control, University of Michigan, May 17-19, 1972. (AFFDL-TR-72-92).

Ninth Annual Conference on Manual Control, Massachusetts Institute of Technology, May 23-25, 1973. (Proceedings published by MIT, no number)

Tenth Annual Conference on Manual Control, Wright-Patterson AFB, Ohio, April 9-11, 1974. (AFIT/AFFDL Report, no number)

Eleventh Annual Conference on Manual Control, NASA-Ames Research Center, May 21-23, 1975. (NASA TM X-62,464)

Twelfth Annual Conference on Manual Control, University of Illinois, May 25-27, 1976 (NASA TM X-73,170)

Thirteenth Annual Conference on Manual Control, Massachusetts Institute of Technology, June 15-17, 1977. (Proceedings published by MIT, no number)

Fourteenth Annual Conference on Manual Control, University of Southern California, April 25-27, 1978 (NASA CP-2060)

Fifteenth Annual Conference on Manual Control, Wright State University, Ohio, March 20-22, 1979. (AFFDL-TR-79,3134)

Sixteenth Annual Conference on Manual Control, Massachusetts Institute of Technology, May 5-7, 1980. (Proceedings published by MIT, no number)

Seventeenth Annual Conference on Manual Control, University of California at Los Angeles, June 16-18, 1981. (JPL Publications 81-95)

Eighteenth Annual Conference on Manual Control, Wright-Patterson AFB, Ohio, June 8-10, 1982. (AFWAL-TR-83-3021)

Nineteenth Annual Conference on Manual Control, Massachusetts Institute of Technolgy, May 23-25, 1983. (MIT publication, no number)

ACKNOWLEDGEMENTS

We would like to thank the following people for the many hours of help that they gave in planning, organizing and running the meeting itself and preparing the proceedings for publication: Michael Bortolussi, Roberta Cortilla, Carl Decker, Sherry Dunbar, Gary Guthart, Robin Karl, Patrick Lester, Dave Marcarian, Ronald Miller, and Richard Rothchild.

Sandra G Hart

Sandra G. Hart, Conference Publications

E. James Hartzell

E. James Hartzell, General Chairman

CONTENTS
VOLUME ONE

FOREWARD i

TIME SERIES MODELING
Chairman: Dr. Greg Zacharias

Biezad, D. & Schmidt, D. K. Time series modeling of human operator dynamics in manual control tasks. 1

Altschul, R. E., Nagel, P. M. & Oliver, F. Statistical time series models of pilot control with applications to instrument discrimination. 41

Bosser, T. Utilization of historic information in an optimisation task. 77

MANUAL CONTROL
Chairman: Mr. Frank George

Jewell, W. & Citurs, K.D. Quantification of cross-coupling and motion feedthrough for multiaxis controllers used in an aircombat flying task. 79

King, M. Six degree of freedom control with each hand? 91

Hess, R. A. & Myers, A. A. A nonlinear filter for compensating for time delays in manual control systems. 93

Schmidt, D. K. & Yuan, P.-J. Model estimation and identification of manual controller objectives in complex tracking tasks. 117

Bekey, G. & Hadaegh, F. Y. Structure errors in system identification. 149

Repperger, D. W. & Levison, W. H. Effects of control stick parameters on human controller response. 157

EFFECTS OF TRANSPORT DELAYS IN SIMULATOR PERFORMANCE
Chairman: Mr. Henry Jex

Jewell, W. & Clement, W.F. A method for measuring the effective throughput time delay in simulated displays involving manual control. 173

Allen, R. W. & DiMarco, R. J. Effects of transport delays on manual control system performance. 185

Sanders, K. E., Anderson, D. C., & Watson, J. H. STOL Simulation requirements for development of integrated flight/propulsion control systems. 201

Privoznik, C. M., Berry, D. T., & Bartoli, A. G. Measurements of pilot time delay as influenced by controller characteristics and vehicle time delays. 209

Kruk, R. V. & Longridge, T. M. Psychophysiological research in development of a fibre-optic helmet mounted display. 221

FLYING QUALITIES
Chairman: Professor Ronald Hess

Heffley, R. K. , Bourne, S.M. & Hindson, W. S. Helicopter pilot performance for discrete-maneuver flight tasks. 223

Onstott, E. D., Warner, J. S. & Hodgkinson, J. Maximum normalized rate as a flying qualities parameter. 233

Sorenson, J. A. & Goka, T. Predictions of cockpit simulator experimental outcome using system models. 259

Hess, R. A. & McNally, B. D. Multi loop manual control of dynamic systems. 281

FAULT DIAGNOSIS/MODELING
Chairman: Dr. William Rouse

Curry, R. & Neu, J. E. A model for the effectiveness of aircraft alerting and warning systems. 299

Gersten, W. M. & Hawkins, J. D. Development and certification of a new stall warning and avoidance system. 301

Laritz, F. & Sheridan, T. Evaluation of fuzzy rulemaking for expert systems for failure detection. 327

Smith, P.J., Giffin, W. C. & Rockwell, T. H. The role of knowledge structure in fault diagnosis. 337

TELEOPERATORS
Chairman: Professor Thomas B. Sheridan

Corker, K. & Bejczy, A. The effect of part-simulation of weightlessness on human control of bilateral teleoperation: Neuromotor considerations. 339

Sheridan, T. B. Review of teleoperator research. 361

Rezek, T. Visual systems for remotely controlled vehicles. 367

SUPERVISORY CONTROL MODELS/TASKS
Chairman: Dr. Azad Madni

Berg, S. & Sheridan, T. B. Measuring workload differences between short-term memory and long-term memory scenarios in a simulated flight environment. 397

Moray, N., Richards, M. & Brophy, C. Visual attention to radar displays. 417

Hart, S. G., Battiste, V. & Lester, P. POPCORN: A supervisory control simulation for workload and performance research. 431

Morris, N. M., Rouse, W. B., Ward, S. L. & Frey, P.R. Psychological issues in online adaptive task allocation. 455

PERCEPTION AND ACTION IN SIMULATOR DISPLAYS
Chairman: Dr. Stanley Roscoe

Levison, W. H. & Warren, R. Use of linear perspective scene cues in a simulated height regulation task. 467

Haines, R. F. Cockpit window edge proximity effects on judgements of horizon vertical displacement. 491

van der Vaart, J. C. & Hosman, R. J. A. W. Mean and random errors of visual roll rate perception from central and peripheral visual displays. 515

McGreevy, M. W. & Ellis, S. R. Direction judgement errors in perspective displays. 531

Stoffregen, T. The interaction of focused attention with flow-field sensitivity. 551

Hosman, R. J. A. W. & van der Vaart, J.C. Accuracy of step response roll magnitude estimation from central and peripheral visual displays and simulator cockpit motion. 559

MODELS OF TARGET ACQUISITION
Chairman: Dr. Carroll Day

Zaleski, M. & Sanderson, P. Hitts' Law? A test of the relationship between information load and movement precision. 575

Jagacinski, R. J., Plamondon, B. D. & Miller, R. A. A production system model of capturing reactive moving targets. 585

Kim, W. S. , Lee, S. H., Hannaford, B., & Stark, L. Inverse modeling to obtain head movement controller signals. 601

Connelly, E. M. A control model: Interpretation of Fitts' Law. 621

Miller, R. A., Messing, L. J. & Jagacinski, R. J. The impact of pictorial display on operator learning and performance. 643

BIODYNAMIC FACTORS

Chairman: Professor John Lyman

Jex-Courter, B. & Jex, H. R. Does McRuer's Law hold for heart rate control via biofeedback display. 663

Winters, J. M. & Stark, L. New uses for sensitivity analysis: How different movement tasks effect limb model parameter sensitivity. 671

Velger, M., Merhav, S. J., & Grunwald, A. J. Suppression of biodynamic interference by adaptive filtering. 699

Repperger, D. W. & Mc Collor Active sticks - A new dimension in controller design. 719

Levison, W. H., McMillan, G. R., & Martin, E. A. Models for the effects of G-seat cueing on roll- axis tracking performance. 735

Hancock, P. A., Carlton, L. G. & Newell, K. M. An analysis of kinetic response variability. 753

Nam, M. -H., & Choi, O. -M. Effects of external loads on human head movement control systems. 761

CONTENTS
VOLUME TWO

EVOKED POTENTIALS

Chairman: Dr. Al Fregly

Donchin, E. The use of ERPs to monitor non-conscious mentation. 1 ✓

Kramer, A. F., Wickens, C. D. & Donchin, E. Performance enhancements under dual task conditions. 21 ✓

Junker, A. M. & Peio, K. J. In search of a visual-cortical describing function. A summary of work in progress. 37 ✓

SUBJECTIVE EVALUATION OF WORKLOAD

Chairman: Mrs. Sandra G. Hart

Gopher, D. Measurement of workload: Physics, psychophysics, and metaphysics. 55 ✓

Vidulich, M. A. & Wickens, C. D. Subjective workload assessment and voluntary control of effort in a tracking task. 57 ✓

Wierwille, W. W., Skipper, J. H. & Rieger, C. A. Decision tree rating scales for workload estimation. Theme and variations. 73 ✓

Miller, R. C. & Hart, S. G. Assessing the subjective workload of directional orientation tasks. 85 ✓

Damos, D. Classification schemes for individual differences in multiple-task performance and subjective estimates of workload. 97 ✓

MENTAL MODELS

Chairman: Professor Neville Moray

Sanderson, P.M. Mental models of invisible logical networks. 105 ✓

Gopher, D., Fussfeld, N., Koenig, W. & Karis, D. The representation of action plans in long term memory. 121 ✓

Rouse, W. B. & Morris, N. M. On looking inside the black box: Prospects and limits in the search for mental models. 123 ✓

Serfaty, D. & Kleinman, D. L. Issues in developing a normative descriptive model for dydactic decision making. 125 ✓

Sheridan, T. B., Roseborough, J., Charney, L, & Mendel, M. Getting mental models and computer models to cooperate. 127 ✓

OTHER ISSUES

Chairman: Professor Larry Stark

Williams, D. H., Simpson, C. A. & Barker, M. A comparative study of alternative controls and displays for the severely physically handicapped. 129 ✓

Stein, A. C., Allen, R. W. & Jex, H. R. A manual control test for the detection and deterrence of impaired drivers. 143 ✓

Agarwal, G. C., Corcos, D. & Gottlieb, G. L. Electromylographic patterns associated with discrete ankle movements. 157 ✓

Kraiss, K. F. & Kuttelwesch, K. H. Color and grey scale in sonar displays. 175 ✓

Wingrove, R. C. Manual control analysis applied to the money supply control task. 181 ✓

CREW FACTORS

Chairman: Cmdr. Kent Hull

Curry, R. E. What pilots like (and don't like) about the new cockpit technology. 199 ✓

Goguen, J. A. , Linde, C. A., & Murphy, M. R. Crew communication as a factor in aviation accidents. 217 ✓

Murphy, M. R., Randle, R. J., Tanner, T. A., Frankel. R. M., Goguen, J. A., & Linde, C. A full mission simulator study of aircrew performance: The measurement of crew coordination and decisionmaking factors and their relationships to flight task performance. 249 ✓

Siesfeld, A., Curley, R. & Calfee, R. Communication on the flight deck. 263 ✓

TRAINING

Chairman: Dr. Robert Hennessy

Poumade, M. L. Determining training device requirements in Army aviation systems. 273 ✓

Mane, A. M., Coles, M. G. H., Karis, D., Strayer, D. & Donchin, E. The design and use of subtasks in part training and their relationship to the whole task. 283 ✓

MULTIPLE TASK PERFORMANCE

Chairman: Professor Stuart Klapp

Casey, E. J., Kramer, A. F. & Wickens, C. D. Representing multidimensional systems using visual displays. 291 ✓

Klapp, S. T. , Kelly, P. A. , Battiste, V. & Dunbar, S. Types of tracking errors induced by concurrent secondary manual task. 299 ✓

Tsang, P. S. & Wickens, C. D. The effects of task structures on time-sharing efficiency and resource allocation optimality. 305 ✓

Soulsby, E. P. On choosing between two probabilistic choice submodels in a dynamic multi task environment. 319 ✓

MEASUREMENT OF WORKLOAD AND PERFORMANCE IN SIMULATION

Chairman: Dr. Anil Phatak

Kim, W., Zangemeister, W. & Stark, L. No fatigue effect on blink rate. 337 ✓

Connelly, E. M. Performance measures for aircraft landings as a function of aircrat dynamics. 349 ✓

Kantowitz, B. H., Hart, S. G., Bortolussi, M. R., Shively, R. J., & Kantowitz, S. C. Measuring pilot workload in a moving-base simulator: II. Building levels of workload. 359 ✓

Milgram, P., van der Wijngaart, R., Veerbeek, H., Fokkerweg, A. Bleeker, O. & Fokker, O. F. Multi-crew model analytic assesment of decision-making demand and landing performance. 373 ✓

Hemmingway, J. C. An experimental evaluation of the Sternberg Task as a workload metric for helicopter flight handling qualities. 397 ✓

Evoked Potentials

N85
14536

UNCLAS

N85 14536

D1

THE USE OF ERPS TO MONITOR NON-CONSCIOUS MENTATION

by Emanuel Donchin

Department of Psychology
University of Illinois

1. Introduction

1.1 The Washington Post Article

On June 3, 1984 the Washington Post carried an article by correspondent Michael Schrage entitled "Technology Could Let Bosses Read Minds." The article continued on the following page under the headline "Privacy Veil May Block Brain Watchers View." In the article Schrage reports that "Researchers in both academia and industry say it is now possible to envision a marketable product that could instantaneously assess whether employees are concentrating on their jobs by analyzing their brain waves as they work." Westinghouse's Research and Development center in Pittsburgh is described as "exploring the use of brain wave analysis - particularly a brain wave known as the P300 - as a means of determining an individual's level of attention and cognitive processing." The manager of the Human Sciences Laboratory in that Center predicts that "within the next 10 years Westinghouse could market a complete system capable of monitoring the mental processing effort of employees as they worked." The article goes on to review the opinions of others who are involved in the study of P300 lending to exchanges with one labor leader and one legal scholar, from Harvard, regarding the degree to which the use of P300 for reading the mind constitutes an "invasion of privacy."

The claims discussed in Schrage's article, and the worries they engender, have appeared frequently, in the past few years, in the public press and in scientific communications. The claims, and the concerns, are triggered by a solid body of evidence accumulated in several laboratories in the two decades since Sutton and his colleagues discovered the P300 (Sutton, Braren, Zubin, & John, 1965). The evidence suggests that the "endogenous" components of the Event Related Brain Potentials (ERP), and in particular the P300, can indeed be used as a tool in the study of cognitive function (Donchin, 1979). Indeed, much of this research has been supported by government agencies specifically in order to determine if it is possible to monitor, by means of the ERP, the operators of complex man-machine systems. The evidence does indicate that the ERP can provide data on aspects of the interaction between operator and task that may otherwise be opaque to monitoring (Donchin, Coles & Gratton, 1984; Kramer, Wickens and Donchin, 1983; Wickens, Kramer, Vanasse, & Donchin, 1983; Isreal, Chesney, Wickens, & Donchin, 1980).

Yet, it must be emphasized that these conclusions have yet to be tested in the crucible of practical applications. In the main, no research has yet been done to translate the laboratory findings into instruments that can be used by design engineers and by system managers. This is due, in part, to budgetary and to practical considerations. However the reluctance to invest in the development of ERP based monitoring may also be due to concerns regarding the appropriateness of using brain-waves to monitor mental activity. It is important therefore to emphasize that the "mind reading" implications of this work are often stated in a misleading and an inflated manner. We can indeed monitor mentation using the ERP. Furthermore, as I will endeavor to show in this paper, the ERPs provide a unique opportunity to monitor non-conscious mentation. Yet, it is not possible, and I believe it will never be possible, to use the ERP to "read minds" in the popular, friday night horror movie, sense of the phrase. My purpose in this lecture is to describe the class of inferences that can be based on ERP data and to emphasize the limits of these inferences. This, however, will not be an exhaustive review of the use of ERPs in Engineering Psychology. Rather, the application, its scope, and its limitations will be illustrated by means of one example. I will precede this example by a brief technical introduction to the methodology used in the study of ERPs.

1.2 Signal Averaging

Event Related Brain Potentials (or ERPs) are extracted from the EEG that can be recorded between a pair of electrodes placed on a person's scalp. The EEG is recorded as a continual fluctuation in voltage. It is the result of the integration of the potential fields generated by a multitude of neuronal ensembles that are active as the brain goes about its business. Within this "ongoing" signal it is possible to distinguish voltage fluctuations that are triggered in neural structures by the occurrence of specific events. This activity, evoked as it is by an external event, is known as the Evoked, or Event Related, Potential. It is but a faint whisper in the polyneural roar of the EEG. However, this whisper tends to follow the same time course whenever its eliciting event occurs. Therefore, when the EEG immediately following an event is examined over an ensemble of records the whispering ERP's are synchronized and their voice, as it were, becomes audible over the conflicting and asynchronous babble of the remaining EEG. Signal averaging is a technique for extracting such faint signals that follow a fixed time course relative to a trigger point. Detailed descriptions of the procedure are readily available (see Halliday, 1982).

The ERP extracted in this fashion takes the form of a series of fluctuations of the voltage between the recording electrodes. The epoch over which an ERP can be observed is on the order of several hundred milliseconds. The ERP is commonly considered to be a sequence of relatively independent components (Donchin, Ritter & McCallum, 1978). The amplitude of the components, their latency and their scalp distribution are the attributes of the ERP that are most commonly used in monitoring brain, and by implication cognitive, activity. Some of the components of the ERP, in particular those that appear within the first 100 msec following the stimulus are manifestation of the transformation, and the communication, of information in the sensory pathways. These "exogenous" components are generally followed by one or more components whose appearance, and patterns of change, vary with the information processing demands placed on the subject. It is these, endogenous, activities that are used in monitoring cognitive activity.

1.3 How Are The ERPs Used in the Study of Cognition?

The monitoring tool to which Schrage's article refers is a record of a voltage change that can be obtained from the scalp of an awake human. These recordings can be obtained rather reliably and our knowledge has advanced to the point that we can predict with relative ease how attributes of these waves will change as a consequence of a variety of experimental manipulations. One readily obtained component is called P300, because it is positive going and its latency is hardly ever less than 300 msec. The P300 is often obtained in the so-called "oddball" paradigm in which a series of stimuli is presented to the subjects; the stimuli can be classified into two categories. If the events in one of the categories occur only rarely, then the rare events elicit an ERP that is characterized by a large, positive going, voltage change that peaks about 300 msec after the eliciting event. This late positivity is the P300 (see Pritchard, 1981; Donchin, 1981; Hillyard & Kutas, 1983 for reviews of the literature).

The P300, and other ERP components, provide an investigator with a set of dependent variables that can be used in the study of cognition. The manner in which these dependent variables relate to various independent variables is well established. However, as yet very little is known about the origin and functional significance of these signals. Evidence regarding the intracranial sources of the potentials is just beginning to emerge. It is likely that the ERPs represent the summation of potential fields associated with individual neurons who, fortuitously, are so oriented that their fields summate. But, as far as we can tell, the summated fields have no functional role in and of themeselves.

The nature of the ERP and the constraints on the interpretation of its physiological significance raises, inevitably, doubts regarding the validity and the utility of inferences made on the basis of these signals. Even though the Press has proven rather sanguine about the promise of ERP in monitoring cognition, the enthusiasm for its use has not proven infectious. Indeed, those who are most in need of techniques for monitoring the operators of complex systems have not been quick to adopt ERPs despite the very strong laboratory evidence for their utility. In part, this reluctance derives from a misunderstanding. It is commonly assumed that to be useful a "physiological" index must be directly involved in the processing activity being monitored. But this, I argue, is not necessarily a valid approach. In fact it is quite possible to conceive of a situation in which "epiphenomenal" indices may prove quite useful.

1.4 The Espionage Metaphor

The process by which the ERPs are utilized, its powers and its limitations, may be clarified by resorting to an analogy. I derive the analogy from electronic snooping. It seems that the design of computers is nurturing a new form of industrial espionage. These high-tech snoops record radiation emitted in the neighborhood of computing devices. It so happens that the structure of electronic data processing devices causes some of the radiation emitted into the environment to be a manifestation of activity internal to the computer. Moreover, it is apparently possible to extract from this radiation, by appropriate computer analysis, useful data about the informational transactions that take place inside the computer. It is as if the information communicated within the computer's functional elements modulates recordable electrical activity in a manner that allows the perspicacious and enterprising spy to "read the mind" of the computer.

It is noteworthy that the activity recorded, and read, by such a spy is not necessarily a meaningful component from the point of view of the computer's information processing activities. The radiation may very well be due to the manner in which the computer was implemented. The availability of these extraneous signals depends on such factors as the choice of components and their packaging, the quality of the shielding. These are factors that are essentially irrelevant to the operation of the computer as an information processing device. Yet, however epiphenomenal, these activities that are "noise" to the computer are very much "signal" to the spy. Provided the technology for extracting the signals exists. Of course, such an indirect method will be used only if more direct methods to access the information of interest are not readily available.

I tend to view the ERPs in much the same way. For reasons having to do with the manner in which the brain is implemented, some of its activities are manifested on the scalp by a voltage change. It is likely that such activity is seen when many neurons are activated in synchrony and the topography with which these neurons are packed is conducive to the summation of their individual fields (Allison, in press). We assume that under the appropriate circumstances and with the appropriate analysis it may be possible to extract from these signals data that help in interpreting the activity of the brain. This is so because, as with the electronic spies, the actual informational transactions that take place within the brain modulate the ERPs, epiphenomenal as they may be, in ways that allow strong inferences about these informational transactions.

Note that, as with the extraneous radiation in the computer, we need not assume that the ERPs in themselves constitute a functional entity in the information processing executed by the brain. All we need to assume is that the intracranial entities that are manifested by the ERP play a role in information processing and that the modulation of the ERP, as the entities it manifests go about their business, is related in a systematic function to the activity of interest. With these assumptions we can observe variations in the ERP and draw inferences regarding the information processing activity. It is these inferences that allow the use of the ERP as a tool in the study of cognitive function.

To illustrate the manner in which the ERPs can be utilized, I will summarize a study by Gratton, Dupree, Coles and Donchin (in preparation) in which variations in the latency of one component of the ERP, the P300, has been used to reveal aspects of processing that accompany the responses of a subject who is performing an oddball task. The key assertion supported by this study is that ERP data can be useful in the examination of processes that are not readily available to introspection. By making the covert overt the ERPs can help in the study of non-conscious processes.

2. The Oddball Paradigm - Using Names

The study discussed here is one in a series of studies employing the Oddball paradigm in which the stimuli were names of individuals commonly used in the American culture. In all cases the series were constructed so that 20% (or, on occasion, 10%) of the names were names of males, (e.g., Jack, John, Eric...). All other names were names commonly associated with females, (e.g., Mary, Vanessa...). On some occasions, the subject was required to count the number of names that fell in one or another category, (a COUNT condition). On other occasions the subject indicated the occurrence of one of the categories by pressing one of two buttons, (a Reaction Time, or RT, condition).

The initial study in this series was reported by Kutas, McCarthy and Donchin (1977). Their subjects were presented with 3 different Oddball series. A "Variable Names" series was constructed from names of males and females as described in the previous paragraph. A "Fixed Names" series included just the names DAVID and NANCY. The third series was a sequence of words, 20% of which were synonyms of "PROD." The subject's task was to press one button in response to such synonyms and to press another button in response to all other words. The rare events in each series elicited a large P300. This was true regardless of the specific task assigned to the subject.

It turned out that the latency of the P300 varied across the 3 conditions. This was particularly noteworthy when the subjects were instructed to be accurate. The shortest latency was observed when the subject discriminated between the two names, David and Nancy. A longer latency is seen when the names vary from trial to trial. The longest latency was associated with the need to decide whether each of a rather disparate list of words is a synonym of PROD. These, and a considerable amount of additional data, lead us to suggest that the latency of the P300 depends on the time required for the evaluation of the stimulus. Subsequent work (McCarthy & Donchin, 1981), demonstrated that the latency of P300 is largely independent of the duration of processes that are involved in the selection and execution of the response. The interesting conclusion from these data has been that the latency of P300 is proportional to the time it takes to categorize the stimuli. If this is the case, the P300 latency may be used as a tool in mental chronometry to measure mental timing uncontaminated by "motor" processes (McCarthy & Donchin, 1983; Donchin, 1981). For studies in which P300 latency is indeed utilized in this fashion see Ford, Mohs, Pfefferbaum and Kopell (1980), Duncan-Johnson and Donchin, (1981), Goodin, Squires, and Starr (1983), Pfefferbaum, Ford, Johnson, Wenegrat, and Kopell (1983), as well as Coles, Gratton, Bashore, Eriksen and Donchin (in preparation).

2.1 The Correlation Between P300 Latency and RT

In a more detailed analysis of the data reported by Kutas et al. (1977), McCarthy and Donchin (1979) examined the relationship between the latency of P300 and the Reaction Time associated with each of the trials in an oddball study using names, sorted according to gender. The analysis capitalized on a filtering technique that allowed the measurement of the latency of P300 on individual trials (Woody, 1967). The principal finding has been that the correlation between P300 latency and RT depends on the strategy adopted by the subjects. When the subjects were instructed to be accurate the correlation between P300 latency and RT was significantly different than zero. On the other hand, when instructed to be fast, the subjects' RTs and P300 latencies were quite uncorrelated. These data supported the suggestion (Donchin, 1979, 1981) that the P300, and the motor response, may each be the culmination of a series of processing activities and that these streams of processing can, in principle, be quite independent of each other.

The P300 latency is assumed to reflect the duration of stimulus evaluation processes. From the evidence on hand it would appear that the processes leading to the invocation of a P300 continue for as long as is required for a full evaluation of the stimulus. The latency of P300 is, therefore, at least as long as the duration of these evaluation processes. The overt responses, on the other hand, may well be released "prematurely" on the basis of limited information. The correlation between Reaction Time and the latency of the P300 will therefore depend on the degree to which the overt responses that define the RT are made contingent on the full evaluation of the stimulus. The more inclined the subject is to respond prematurely, the poorer the correlation between the latency of the P300 and the RT.

2.2 The P300 On Error Trials

One striking aspect of the data acquired by McCarthy and Donchin (1979) was observed when the trials on which subjects made errors. These were trials on which the subject responded to a rare event as if it was frequent. That is, even though a Male name appeared on the screen, the subject pressed the button associated with Female names. There were but a few such trials in the study reported by McCarthy and Donchin (1979). However, in virtually all these trials the pattern was the same - the Reaction Times were relatively short and the P300 latency was relatively long. It was as if on these trials the subjects first acted and than thought! As the number of error trials was small, we replicated the experiment presenting the subjects with many more trials and pressing even harder for fast responses. A partial report on these data can be seen in McCarthy (in press). In 10 out of the 11 subjects the pattern obtained was identical. Errors of commission, "fast Guesses," were associated with very short RTs and relatively long P300 latencies. McCarthy and Donchin (1979) suggested that whenever an error was detected on any given trial, the invocation of the P300 was delayed. The delay was required, presumably, to allow further processing of the trial's data. This interpretation exemplifies the manner in which observations of the P300 lead to inferences regarding an internal process even though these processes may not be readily observable by conventional means.

2.3 Puzzles for Present Experiment

Even though the increase in the latency of the P300 was quite evident in the data obtained by McCarthy and Donchin (1979) it was not sufficient to support the conclusion that this delay is due to extended processing consequent on an internal, not necessarily conscious, recognition of the commission of the error. Several alternate explanations can be invoked. Two of these difficulties are summarized here.

2.3.1 New Component?

One of the major difficulties presented by ERP data is associated with the definition and the proper identification of components of the ERP. For example, each of the positive going peaks observed by Kutas et al. (1977) in the ERPs elicited by the three series has been labeled "P300" even though the peaks differ in latency by as much of 100 msec. What leads us to believe that these three peaks are indeed instances of a component whose latency is shifted by the duration of the processing precedes its invocation? How do we know that the peaks with the longer latencies are not entirely new components that are elicited by the presentation of a word, or by the search of a synonym. The issue is generally resolved on the basis of the similarity of wave shapes, on the scalp distribution of the potentials and on the manner in which they respond to experimental manipulations (Donchin, et al., 1978). There remains the possibility that delayed peaks that are recorded in association with error trials are different components rather than a delayed P300.

2.3.2 Response Related Factors

Another interpretation of these data is based on the fact that on all these error trials the subject responded rather fast to the stimulus. In other words, these are clearly trials on which a variety of factors are injected into the stream of processing. How do we know that it is the recognition of the error, rather than the fact that a very fast response was emitted on the trial that accounts for the delay? A different, but related possibility is that it is not that P300 is delayed on error trials, but rather that errors may be more likely on trials on which P300 latency is long.

2.3.3 The Need For an Additional Study

The controversy surrounding the interpretation of the ERPs recorded on error trials touches on some of the key issues in the interpretation of the ERP. The manner in which such controversies arise, and the action that is needed to resolve the issue, must be understood if these data are to be used in the, so-called, "real" world. Any monitoring system that utilizes ERPs in the manner described by the Washington Post article will, in one way or another, acquire data much like those described above. Essentially the data analysis, however sophisticated, boils down to a comparison of the amplitudes of waveform features obtained at different sites on the same occasion or features that were obtained from the same site on different occasions. Whenever such a comparison is made it is critical to assure that one compares features of the same object. If it is possible to mistake one component for another, then shifts in latency or in amplitude that are assumed to reflect shifts in the allocation of attention may in fact reflect an altogether different process. Such a confusion will frustrate any attempt to utilize the ERPs, regardless if the use is made in a laboratory or an industrial environment.

In the present case, the claim that is in need of evaluation is that the P300 reveals, through modulations of its latency, the activation of an internal, mental, process that is invoked as a consequence of the recognition that an error has occurred. If we can be sure that the peak with the longer latency is indeed a delayed P300 rather than a new component, and if we can be sure that the delay is indeed due to the occurrence of the error rather than to such factors as the speed of the response associated with the movement, than the P300 is indeed revealing in a unique fashion aspects of the information processing system. To resolve some of the doubts that remained regarding the ERPs elicited on error trials we replicated, and extended, the study reported by McCarthy and Donchin (1979).

3. A Study of P300 Latency on Error Trials

Thus, we have again presented subjects with a series of names. In one series the names appeared with unequal probability, names of Females appearing frequently, P(female)=.80. In another experimental condition the two categories appeared with equal probability. These two probability conditions were crossed with two performance regimes. In one the subject was instructed to respond as fast as possible. In the other regime the subject was told to be as accurate as he could. From each of the 7 subjects we obtained 800 trials in each of the conditions.

3.1 Design

Procedure. The subject was positioned in front of a PLATO terminal with the fingers of each hand resting around a 2" diameter bar of a dynomometer. The choice-reaction-time task required a sharp squeeze and release of the bar from one hand in response to male names appearing on the screen and a squeeze and release from the other hand in response to female names. Names were presented one at a time in the center of the screen for 200 msec with a 2000 msec interstimulus interval. A list including 10 male names and 10 female names was used to generate the series. The four to seven character names were chosen for their familiarity and for the certainty of their gender.

Subjects were shown the names in blocks of 100 trials. Blocks were made up of either 80 females and 20 males or 50 of each. Also, subjects were instructed to respond as quickly as possible or as quickly as possible without making errors. The two conditions, (1) the relative probability of male and female names and (2) the instruction set (speed or accuracy), were factorially combined, resulting in eight experimental cells. Eight hundred trials were run in each cell, with half the trials run during one session and the remaining half run during a second session. During each session, four blocks of 100 trials were run for one experimental cell at a time. The order of conditions was counterbalanced across subjects in a latin square design, and the order of conditions run during the first session was reversed for the second session. Also, the relationship between the class of stimuli (male or female names) and the responding hand (left or right) was counterbalanced across subjects.

In addition to response time, EEG was recorded by Ag-Ag Cl electrodes at Fz, Cz, Pz, C1, and C2 placed according to the 10-20 system and referred to linked mastoids. EOG was recorded for purposes of subtracting out ocular artifact from EEG, with a Beckman electrode placed above and to the right of the right eye. EMG was recorded by two Beckman electrodes placed one half an inch apart, one third of the distance on the diagonal between the elbow and the outer wrist when palm up. Analog to digital conversion occurred for 1200 msec which consisted of 100 msec of baseline before each stimulus name and 2200 msec from the movement of presentation.

3.2 Results

A detailed presentation of the rather large amount of data, and the numerous analyses of these data will be given in Gratton, et al. (in preparation). Here, I shall summarize some of the results focusing on the data obtained when the Male names where rare an the subject was urged to be fast, (the "speed" condition). I will not present here the statistical analyses that support my various assertions. Again, these are presented with some detail by Gratton, et al. (in preparation). The reader can rest assured that all statements made here are backed by adequate analyses.

3.2.1 Reaction Time Data

3.2.1.1 Histograms for Individual Subjects

The pattern of Reaction Times was consistent. Subjects respond with virtually no errors to Female names. They do so rather fast. That is, the RTs associated with female names tend to be short and the number of errors, that is presses on the Male button in response to a Female name, is miniscule. The pattern for Male names is quite different. Correct responses to Male names are rare and, when given, they are given slowly. On the other hand, it is clear that on most trials on which a Male name is the stimulus the subject presses the "Female" button. Moreover, the RT on these trials tends to be quite short. The RT in this case is in fact quite similar to the RT associated with the correct Female name.

The data indicate that the subjects' responses differed according to the button they pressed, or the hand they were using. The Male button was pressed solely in response to the appearance of Male names. The speed with which these responses were made was always slower than was the speed of response on the Female button. It is plausible to assume that the subjects were primed to respond with the hand that was called upon to respond most frequently. This "response bias" caused the subject to respond on many a trial to the Male name with the response on the Female button. It is striking that the distribution of the RTs for these fast responses is rather independent of the eliciting stimulus. Pressing, correctly, for a Female name and committing a "fast guess," by pressing the same button in response to a Male name are indistinguishable as responses, at least as far as the shape of the distribution is concerned.

3.2.2 ERP Data

While the correct overt response is indistinguishable from the overt erroneous response the processing associated with the two classes of responses is likely to be quite different. There is considerable evidence that fast guesses, and other errors, are monitored and processed by the subject and that the performance on subsequent trials is affected by such processing. This error processing need not call on the subject's awareness. The error may be processed, and its consequences integrated into the response stream, whether or not the subject is conscious of the error. Indeed, the existence of error-related processing has heretofore been inferred from variations in the performance on trials that follow the error. An examination of the ERPs acquired by Gratton, et al. reveals that some intracranial processing entity is affected by the occurrence of an error.

Support for this claim is provided by examination of the ERPs elicited by names of Males and of Females. The EEG activity was sorted so that the ERPs associated with correctly identified and mis-identified Male names are plotted separately, as are the responses to Female names. The data were also sorted according to the speed of the response. The bottom panel plots the data from the fastest responses, each successive panel represents slower responses. The data are clear. The ERPs elicited by the missed Male names and by the Female names are quite different in pattern. The Male names elicit a substantial P300, the Female names barely do. Thus, the homogeneity of the motor responses obscures a difference between the activity of whatever intracranial system is manifested by the P300. As the response topography of the Male and Female responses appears to be quite similar, it is difficult to attribute the delay in the latency of names to the P300 to the speed with which the subjects respond on the error trials. The speed of the response on a Female trial is equal to the speed of the response on the incorrect Male trial.

There is also a patent difference between the ERPs elicited by Male trials that were correctly identified and those that were missed. The peak positivity on the error trials is delayed by almost 100 msec. This finding corroborates the reports by Kutas, et al. (1977). A detailed analysis of the distribution of the component supports the identification of the delayed component as the P300 (see Gratton, et al., in preparation). Thus, we confirm the paradoxical relationship between the RT and the latency of P300. The relatively short RT's associated with the incorrect trials are accompanied by a P300 with a long latency. Conversely, when the RT is relatively long, as it is on the correct trials, the P300 latency is short. It is important to note that this pattern of results holds for all the conditions used in this study. Error trials were associated with the longer latency P300s when the probability of names in the two categories was equal. Similarly, the result held when subjects tried for accuracy. Moreover, the pattern was maintained even when the data were sorted according to the speed of the response. That is, when trials are classified into bins according to the RT on each trial, then within each bin the error trials are associated with longer latency P300.

3.3 Interpretation

It seems, therefore, that it would be prudent to accept the empirical assertion that the P300 tends to have a substantially longer latency on trials on which the subject pressed the wrong button. How can we interpret such an observation? What, if anything, does it tell us about the mental activities that take place as the subject is performing the assigned task? The empirical statement, by itself, can support the conclusion that there is a difference of some sort between processing activities accompanying correct and incorrect trials. But, establishing the existence of such a difference is not a particularly satisfying enterprise. In the first place it is not all that surprising that such a difference is observed. Moreover, the existence of such differences has been established quite persuasively by means of the classical methods of Cognitive Psychology. What do we gain, how do we augment the available knowledge, by adding the ERP to our armamentarium?

3.3.1 The ERP and Non-Conscious Mentation

It would seem that one of the principal values of the ERPs is that they allow observation of processes that do not have obvious representations in awareness. That such processes exist goes almost without saying. We are not aware, and most probably can not be aware, of most of the internal information processing activities that yield as a consequence the contents of awareness. Consider Speech. By and large we are aware of the content of our discourse. We know what we say, we may know why we want to say what we say and we know the purpose underlying our words. These all are the contents of consciousness. Yet, we are at the same time entirely unaware of the nature of the process used to select our vocabulary, or sort out these words into proper grammatical sentences. Even when we consciously search for a word, we are blissfully unaware of the manner in which our mental gears grind as the word is searched for. When candidate words are dredged, we know immediately - we are fully "aware" of - the degree to which that word is, or is not, a suitable choice. But, if we know it is not the correct word, how come we cannot find the proper word? These processes, and much more that is of interest to the cognitive scientist, takes place well outside consciousness.

It is in fact these non-conscious activities that are the principal focus of interest to Psychologist. True, as persons we are principally interested in that of which we are aware. But, as Psychologists we are interested in the processess underlying the observed behavior. We would like to understand how memory is organized and how information in memory is searched and is retrieved. We would like to know how sensory information is integrated into the percepts of objects and how the speech stream is scanned into words whose meaning is extracted even as all their related associations are activated. These are the psychological operations whose elucidation is the goal of Cognitive Psychology. As these are largely non-conscious the Science is based on inferences from observations on the pattern of overt behavior. Alternately we depend on self-reports, a rich but occasionally flawed record. It seems that, at least to a limited extent, ERP components allow us to monitor directly the intensity and the latency of some of these processes (Johnson & Donchin, 1978; Johnson & Donchin, 1982; Donchin, McCarthy, Kutas, & Ritter, 1983).

3.3.2 The Research Design

But, even if one grants that the ERP is a manifestation of brain activity which implements an interesting mental operation, and hence by implication the ERP can be considered a manifestation of such mental operations, how does one determine the nature of the specific operations associated with a specific component. Clearly the degree to which the P300 or any other component could be used for monitoring operators depends on the degree to which the functional significance of the component is known. In the specific case we are discussing here we need to have a theory regarding of the functional significance of the P300 so that a framework is available for assessing the implication of its increased latency.

How does one go about elucidating the functional significance of an ERP component? In my view a three fold process is required (see Donchin, 1981; Donchin & Bashore, in press; Donchin, et al., 1984). The entire process is guided by a view that sees an ERP component as the manifestation of an intracranial processor which implements some information processing operator. This statement raises some complex philosophical issues (see Donchin & Bashore, in press). However, in its simplest form the relation between the ERP and mentation is viewed in much the same form as are the radio emissions discussed in Section 1.4. The principal implication of this view is that theories regarding the functional significance of the ERP are best developed within some comprehensive model of information processing. The hypothesis regarding the component's function will be stated by identifying a processing element within the general model. Such an element is defined in terms of the transformations it performs on its input. A theory of the P300 than asserts that the component's appearance indicates that this particular operation has been invoked. The component's latency is a measure of the duration of processes whose occurrence must precede the invocation of the processor. The amplitude of the component is taken as a measure of the intensity with which the critical operation has been performed. Many assumptions are implicit in this description of theory building in Cognitive Psychophysiology. Some are more tenuous than others. Thus, inferences about the latency of a component are fairly straightforward. On the other hand, the interpretation of the amplitude as a measure of the utilization of the component (Donchin, Kubovy, Kutas, Johnson, & Herning, 1973) is based largely on faith, on the plausibility of the assumption and on the fact that this is as good a working hypothesis as we can muster.

3.3.3 The Need for Theory Testing

A theory of the P300 must begin with an enumeration of what I have called the antecedent conditions of the component (Donchin, 1981). In effect, the bulk of the research on P300, including the study described in detail in this lecture, has been concerned with the enumeration of these antecedent conditions. This search yields an ensemble of statements that describe the conditions under which the P300 is elicited. There is also a need to determine the functional relationship between variations in many aspects of the eliciting situation and attributes of the P300. Much effort has been invested in determining the factors that control the amplitude of the P300, its latency and the variation in its scalp distribution. Such data have accumulated in the last two decades to an extent that permits a rather precise enumeration of the antecedents of the P300.

3.3.4 The Antecedents of the P300

The list is familiar (Hillyard & Kutas, 1983; Pritchard, 1981). The P300 is elicited by rare, task relevant, events. If task relevance is held constant than the amplitude of P300 is inversely proportional to the subjective probability of the eliciting event. If subjective probability is held constant than P300 amplitude is determined by the extent to which the task with which the P300 is associated is at the focus of the subject's attention. This indeed is the basis for the use of P300 as a measure of Workload (Isreal, et al., 1980; Kramer, et al., 1983; Donchin, Kramer & Wickens, 1982). It is also clear that while the rarity of the eliciting event can play an important role in the elicitation of the P300, rarity is neither a sufficient, nor a necessary, condition. Studies of P300 elicited when subjects are assigned dual tasks indicate that P300 is a manifestation of processes associated with perceptual, categorization, activities. In addition evidence has been presented that the amplitude of P300 is inversely proportional to the degree to which an earlier representation of the stimulus has decayed (Squires, Wickens, Squires & Donchin, 1976).

With these ensemble of antecedents on hand one can proceed to the next two stages of the theory building process. These data, if sufficiently complete can lead to a model of the P300 couched in the terms we required above. That is, a statement need be made that assigns a function to the P300. The statement represents an integration and an interpretation of all that we know about the P300's antecedent conditions. To be useful it is not sufficient for this model to be merely a plausible summary of the available data. Rather, it should serve as the basis for the third, the theory testing, phase. In that last phase predictions that are derived from the hypothesis we entertain regarding the component's function need be tested. Such predictions take the form of statements about the consequences of the P300.

As I argued elsewhere (Donchin, 1981), if the P300 is a manifestation of a processing entity, a subroutine if you will, than it must have outputs that feed into subsequent, or parallel, stages of the information processor. If the amplitude of the component is proportional to the intensity of its activation, than its activity will affect subsequent processing stages in a manner that is related to the amplitude of P300. In other words, it must have consequences. If we believe we know its function, we ought to be able to predict these consequences. It is in the generation and the testing of such hypotheses that theories regarding the P300 are tested.

4. A Hypothesis Regarding the P300

The specific hypothesis that currently serves as a guide for the work my colleagues and I are conducting at the Cognitive Psychophysiology Laboratory at the University of Illinois views the process manifested by the P300 as an instrument in the service of the operation of Working Memory. By this term we refer to the ensemble of representations that are, at any time, in a state of higher availability. The membership in this ensemble is continually changing as the needs of the moment change. For any given task, some new representations may be needed, while others (remaining from previous tasks) must be discarded. The process is dynamic and requires, one should assume, a considerable amount of housekeeping. There must be an ongoing process of context evaluation and context updating. I have argued that the P300 is a manifestation of a processing entity that is utilized while such context updating, or memory management, takes place.

Whether this model will ultimately prove to be a good approximation to the truth remains to be seen. However, it does satisfy the criteria for a model in that specific predictions can be derived from that model regarding the consequences of the P300. For example, Klein, Coles and Donchin (1984) have shown that people with perfect pitch process phonic probes without emitting a P300. That this would be the case was predicted on the basis of the context updating hypothesis. Karis, Fabiani and Donchin (1984) have shown than the amplitude of the P300 elicited by a stimulus in a study of the von Restorff effect predicts whether or not the stimulus will be recalled.

4.1 The Delayed P300 on Error Trials--An Interpretation

If the process manifested by the P300 performs a function that is necessary for the maintenance of the model of the environment in Working Memory than it may be suggested that it is not invoked until the data needed for determining the needed changes is available. We propose that the delay in the P300 on error trials is inserted as the error is recognized by the system because there is a need for further processing before the book can be closed on the trial. Note, than in our view the P300 process is invoked in order to serve the needs of action on future trials. Thus, the elicitation of P300 on when the rare stimulus appears may be associated with the resetting of the system to accommodate responses to the rare events. After all, the subject is clearly biased to emit the frequent response at the slightest provocation. One assumes that these responses are emitted as soon as the appearance of a stimulus is detected. As processing of the stimulus continues, after the response has been made, the name is properly encoded. The conflict between the category of the name and the response forces on the system additional processing. The additional time required for this processing is the delay we observe in the P300.

We are fairly confident that the delay in P300 on error trials is indeed associated with the recognition of the error. Though we emphasize that we are not implying that this is a conscious, intentional, delay. Other plausible alternatives have been considered and have been ruled out, (Gratton, et al., in preparation). The proposal is plausible. However, the plausibility does not provide adequate support for the theory. The critical test, again, is the ability to derive from our interpretations of the delay specific predictions. In this case, the proposal that the process manifested by P300 serves the responses made by the subject on future trials suggests that there ought to be a relationship between the amplitude of the P300 elicited on error trials and performance on succeeding trials. We conducted two such tests to evaluate the validity of this view.

4.2 The Amplitude of P300 on Error Trials And Its Consequences

If subjects err because they are biased to respond to the frequent event than one consequence of the recognition of an error would be an attempt to shift the bias away from the activation of the frequently pressed button. The shift would be in the direction of the response to the rare event. Such a shift should be accompanied by an increased probability that a response will be given on the "male" button to male name. If the P300 is an index of the degree to which readjustments of the system's model of the environment than, the larger the P300 the large we would expect the shift to be. We examined therefore the subject's responses on all trials in which a Male name was presented. It turns out that the larger the P300 elicited on an error trial, the more likely is the subject to be correct in the response the next time a male name is presented regardless of the number of female names that have appeared in the interim. It would appear that subjects indeed modulate their response bias when an error is discovered. More important is the observation that the degree to which this shift in strategy takes place is indexed by the P300. These data strongly support the proposition that the amplitude of the P300 reflects the intensity with which a context-updating process has operated.

That there is indeed a shift in the bias is supported by the analysis of the Reaction Times associated with the presentation of Female names that occurred immediately after an erroneous response was made to a male name. If the subject is indeed shifting response bias in the direction of Male names, we expect the responses to female names to be slowed down in the trials immediately following missed Male names. This increase in RT should be proportional to the amplitude of the P300 elicited on the error trial. This, is precisely what we found. The larger the P300 elicited on a given trial the slower is the response to the immediately following female names.

It is interesting that the latency of the P300, delayed as it may be, does not predict the response on subsequent trials. But, than, this should come as no surprise. The latency is index of the duration of the processes preceding the invocation of the P300. Thus, it is not directly related to the process which in fact updates the context. The latency should therefore should therefore have no effect on the subject's model of the environment. And indeed, we could detect no relationship between P300 latency and subsequent performance.

5. Conclusions

5.1 The Implications for Monitoring

The nature of the information on mentation that can be gleaned from ERPs is illustrated by the data I have just described. The study is quite typical in the evidence it yielded and in the complexity of the procedures required to interpret the evidence. How likely is it that devices for measuring the P300 will appear, let alone proliferate, in the work place in the coming decades? It seems clear that the ERPs do provide information that is not otherwise available. However, it should be equally clear that the language with which the ERPs speak is arcane. The significance of the presence, or absence, of a P300 and the interpretation of modulations of its amplitude and latency can be assessed only within the framework of a careful analysis of the circumstances. The amplitude of P300 can increase, or decrease, for a large number of different reasons. In a carefully structured situation the interpretation, to the trained and skilled investigator, is not too difficult. But, it is unlikely that it would be possible to attach a machine that would yield a simple, universal, situation-independent, number that can be used by a manager, a designer, or even the operator to make intelligent on-line decisions.

Of course, it is not my intention to suggest here that the efforts to develop the ERP as a tool for the Engineering Psychologist were wasted. I do believe that the ERP is a unique and valuable tool. However, it must be realized that, as is true for any tool, it is best used within the constraints of its nature and it better be applied within contexts that justify its usage. The available literature defines the nature of the information about an operator that can be extracted from the ERP. Whether this information is of utility in any given situation depends on the degree to which the information be utilized. If, for example, a man-machine system is not adaptive then it is entirely wasteful to provide it with information on the shifts in the operator's level of attention. The very same information may be extremely valuable, and well worth the cost of data-acquisition, if the system within which it is obtained is capable of adjusting to the operator's level of attention. In other words, the Psychophysiologist can point the availability of the information and define the methods by which it can be acquired. It is for the engineer and system designer to determine if this information can improve system performance at a reasonable cost.

One, of course, cannot be sanguine about the matter. If Polygraphy (lie-detection) can be used as a case in point, we must admit that when a technology that is capable of commercial exploitation becomes available the potent mix of the unscrupulous and the gullible may generate a vast industry. Polygraphy, like ERP research, utilizes a reliable phenomenon. It capitalizes on the fact that emotional changes are manifested by a class of recordable bodily changes. The interpretation of these changes in any given situation requires skill and a very careful analysis of the psychological structure of the situation. It may, in very carefully designed tests, in the hands of well-trained, experienced, Psychophysiologists yield valuable information about the veracity of a witness. To move from this to the application of the polygraph in personnel offices to screen job applicants is bizarre indeed. I dearly hope that we shall not see in the near future the appearance of ERPgraphers, wielding Signal Averagers, assessing workers' productivity to the joy of gullible corporate managers.

The need to guard against the avaricious and the naive should not obscure the vast possibilities opened by Cognitive Psychophysiology for a better understanding of human performance, and for monitoring operators in useful ways. The P300, and the other ERP components, clearly provide useful data. Our knowledge of these signals is still in its earliest stages. I am confident, however, that the range of useful information that can be extracted from the ERP will be extended in the coming decades. There is already sufficient data to justify the incorporation of ERP measures in the design phase of complex systems. The closed-loop application that comes to mind when we consider monitoring an operator may be a thing of the remote future. However, the P300 can be of considerable use to designers who need to evaluate several competing systems in terms of the effectiveness with which operators can use the systems. The development effort, to my mind, should be devoted largely to the utilization of this valuable window on the mind in the design, rather than during the actual use, of Person-Machine systems.

References

Allison, T. Recording and interpreting event-related potentials. In Donchin, E. (Ed.), Cognitive Psychophysiology Proceedings of Carmel I. Erlbaum, in press.

Coles, M.G.H., Gratton, G., Bashore, T.R., Eriksen, C.W., & Donchin, E. A psychophysiological approach to the continuous flow model of cognitive processes, in preparation.

Donchin, E. Event-related brain potentials: A tool in the study of human information processing. In H. Begleiter (Ed.), Evoked potentials and behavior. New York: Plenum Press, 1979, pp. 13-75.

Donchin, E. Surprise! . . . Surprise? Psychophysiology, 1981, 18, 493-513.

Donchin, E., & Bashore, T. Clinical versus psychophysiological paradigms in the study of event-related brain potentials. In S. Harnad (Ed.) The behavioral and brain sciences, in press.

Donchin, E., Coles, M.G.H., & Gratton, G. Cognitive psychophysiology and preparatory processes: A case study. In Kornblum, S.N. and Requin, J. (Eds.), Preparatory States and Processes, Hillsdale, NJ: Erlbaum Associates, 1984, 155-178.

Donchin, E., Kramer, A., & Wickens, C. Probing the cognitive infrastructure with event- related brain potentials. In Frazier, M.C., & Crowbee, R.B. (Eds.). Proceedings of the workshop on flight testing to identify pilot workload and pilot dynamics, AFFTC-JR-82-5, Edwards Air Force Base, 1982, 371-387.

Donchin, E., Kubovy, M., Kutas, M., Johnson, R., Jr., & Herning, R.I. Graded changes in evoked response (P300) amplitude as a function of cognitive activity. Perception and Psychophysics, 1973, 14, 319-324.

Donchin, E., McCarthy, G., Kutas, M., & Ritter, W. Event Related Brain Potentials in the study of consciousness. In Davidson, R. Schwartz, G. and Shapiro, D. (Eds) Consciousness and Self Regulation, Vol 3, Plenum Press, 1983, pp 81-121.

Donchin, E., Ritter, W., & McCallum, C. Cognitive psychophysiology: The endogenous components of the ERP. In E. Callaway, P. Tueting, & S. Koslow (Eds.), Brain event-related potentials in man. New York: Academic Press, 1978, pp. 349-441.

Duncan-Johnson, C., & Donchin, E. The relation of P300 latency to reaction time as function of expectancy. In H. H. Kornhuber and L. Deecke (Eds.), Motivation, motor and sensory processes of the brain: Electrical potentials, behavior and clinical use. Progress in Brain Research. Amsterdam: Elsevier-North Holland, 1981, pp. 717-722.

Ford, J.M., Mohs, R.C., Pfefferbaum, A., & Kopell, B.S. On the utility of P300 and RT for studying cognitive processes. In Kornhuber, H.H. & Deecke, L. (Eds.), Motivation, motor and sensory processes of the brain: Electrical potentials, behavior and clinical use. Progress in brain research, Vol. 54, 1980, Elsevier/North Holland: Amsterdam.

Goodin, D.S., Squires, K.C., & Starr, A. Variations in early and late event-related components of the auditory evoked potential with task difficulty. Electroencephalography and Clinical Neurophysiology, 1983, 55, 680-686.

Gratton, G., Dupree, D., Coles, M.G. H., & Donchin, E. An electrophysiological manifestation of an error processing routine. In preparation.

Halliday, A.M. (Ed) Evoked Potentials in Clinical Testing. Churchill Livingstone: London, 1982, pp 575.

Hillyard, S., & Kutas, M. Electrophysiology and cognitive processing. Annual Review of Psychology, 1983, 34, 33-61.

Isreal, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. Psychophysiology, 1980, 17, 259-273.

Johnson, R. E., Jr., & Donchin, E. On how P300 amplitude varies with the utility of the eliciting stimuli. Electroencephalography & Clinical Neurophysiology, 1978, 44, 424-437.

Johnson, R. Jr. & Donchin, E. Sequential expectancies and decision making in a changing environment: An electrophysiological approach. Psychophysiology, 1982, 19, 183-200.

Karis, D., Fabiani, M., & Donchin, E. P300 and memory: Individual differences in the von Restorff effect. Cognitive Psychology, 1984, 16, 177-216.

Klein, M., Coles, M.G.H., & Donchin, E. People with absolute pitch process tones without producing a P300. Science, 1984, 223, 1306-1309.

Kramer, A.F., Wickens, C.D. & Donchin, E. Analysis of the processing requirements of a complex perceptual-motor task. Human Factors, 1983, 25(6), 597-621.

Kutas, M., McCarthy, G., & Donchin, E. Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time. Science, 1977, 197, 792-795.

McCarthy, G. Stimulus evaluation time and P300 latency. In E. Donchin (Ed.), Cognitive Psychophysiology Proceedings of Carmel I, Erlbaum, in press.

McCarthy, G. & Donchin, E. Event-related potentials: Manifestations of cognitive activity. In F. Hoffmeister and C. Muller (Eds.), Bayer-Symposium VII, Brain Function in Old Age. New York: Springer-Verlag, 1979, pp. 318-335.

McCarthy, G., & Donchin, E. A metric for thought: A comparison of P300 latency and reaction time. Science, 1981, 211, 77-80.

McCarthy, G., & Donchin, E. Chronometric analyses of human information processing. In A.W.K. Gaillard and W. Ritter (Eds.), Tutorials in event-related potential research: Endogenous components. Advances in Psychology, Vol. 10, G.E. Stelmach and P.A. Voon (Eds.). Amsterdam: North Holland Publishing Co., 1983, pp. 258-268.

Pfefferbaum, A., Ford, J.M., Johnson, R., Wenegrat, B., & Kopell, B.S. Manipulation of P300 latency: Speed vs. accuracy instructions. Electroencephalography and Clinical Neurophysiology, 1983, 55, 188-197.

Pritchard, W.S. The Psychophysiology of P300. Psychological Bulletin, 1981, 89, 506-540

Squires, K. C., Wickens, C., Squires, N. K., & Donchin, E. The effect of stimulus sequence on the waveform of the cortical event-related potential. Science, 1976, 193, 1142-1146.

Sutton, S., Braren, M., Zubin, J., & John, E.R. Evoked-potential correlates of stimulus uncertainty. Science, 1965, 150, 1187-1188.

Wickens, C., Kramer, A., Vanasse, L., & Donchin, E. The performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information processing resources. Science, 1983, 221, 1080-1082.

Woody, C.D. Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. Medical and Biological Engineering, 1967, 5, 539-553

N85
14537

UNCLAS

D2

N85 14537

Performance Enhancements Under Dual-Task Conditions

Arthur F. Kramer, Christopher D. Wickens and Emanuel Donchin
Cognitive Psychophysiology Laboratory
University of Illinois
Champaign, Illinois

SUMMARY

Research on dual-task performance has been concerned with delineating the antecedent conditions which lead to dual-task decrements. Capacity models of attention, which propose that a hypothetical resource structure underlies performance, have been employed as predictive devices. These models predict that tasks which require different processing resources can be more successfully time shared than tasks which require common resources. We suggest that dual-task decrements can be avoided even when the same resources are required by both tasks, by designing the tasks so that the processing demands can be integrated. The conditions under which such dual-task integrality can be fostered were assessed in a study in which we manipulated three factors likely to influence the integrality between tasks: inter-task redundancy, the physical proximity of tasks and the task relevant objects. The resource structure associated with these integrated dual-task pairs is inferred from changes in the amplitude of the P300 component of the Event-Related Brain Potential (ERP).

Twelve subjects participated in three experimental sessions in which they performed both single and dual-tasks. The primary task was a pursuit step tracking task. The secondary tasks required the discrimination between different intensities or different spatial positions of a stimulus.

Task pairs which required the processing of different attributes of the same object resulted in better performance than task pairs which required the processing of different objects. Inter-task redundancy, the physical proximity of task related stimuli and processing priorities also affected the performance of dual-task pairs. The results are discussed in terms of a model of dual-task integrality.

INTRODUCTION

The concurrent processing of information relevant to several tasks has interested psychologists from the early writings of James to contemporary investigations of dual-task performance in complex, operational environments. Substantial theoretical and empirical effort has been expended in mapping the conditions under which the demands imposed by tasks performed concurrently interact so that performance on one, or both, tasks degrades. Wickens (1980) proposed a Multiple Resource Model according to which processing resources may be represented by three dimensions: stages of processing, modalities of processing and codes of processing. The extent of dual-task interference is predicted on the basis of the overlap of processing resources. Tasks which require separate processing resources will be more sucessfully time shared than tasks which require common processing resources. This theoretical conceptualization of the processing structure of

PRECEDING PAGE BLANK NOT FILMED

dual-task performance has received considerable empirical support. Tasks which require processing resources from the same modalities, codes or stages of processing result in larger performance decrements than tasks which require resources from different structures (Trumbo, Noble & Swink, 1967; Isreal, 1980; Alwitt, 1981). When two concurrently performed tasks require entirely different processing resources, increasing the difficulty of one task fails to have an effect on the performance of the other task (North, 1977; Wickens & Kessel, 1979).

Although the contemporary resource models provide empirically verifiable hypotheses concerning the decremental effects of dual-task performance, they do not address the issue of the integration of the processing of one task with the processing of another task. Under some dual-task conditions, the processing of one of the tasks may prove benefical to the processing of the other task (Wickens & Boles, 1983). For example, subjects may be required to perform concurrently two separate tasks. One task may require tracking a target with a cursor along a single axis on a CRT. The other task may necessitate a discrimination between flashes of different intensities. What if an event in one task now predicts a change in the second task with some degree of certainty? Using the example illustrated above, the spatial position of the tracking target may predict the brightness of the secondary task stimulus. The specific tasks have not changed as a function of the change in the inter-task redundancy. However, the processing of the spatial changes in the tracking task may now benefit the performance of the intensity discrimination task. This effect in which the inter-task redundancy results in performance enhancements has been termed a "concurrence benefit" (Navon & Gopher, 1979). Thus, it is not the tasks that become integrated but instead the processing of the tasks. The present study will investigate the conditions under which performance on one task results in enhancements in performance of a second, concurrently performed task.

The resource model employed in the description of dual-task decrements can also be applied to the examination of integrality between tasks. The phenomenon of dual-task integrality occurs when two separate, but concurrently performed tasks can be processed within the same resource framework. In most dual-task cases, increasing the difficulty of one task is assumed to consume resources which would normally be employed in the processing of the other task. Thus, the allocation of resources between the two tasks is assumed to be reciprocal. However, under conditions of dual-task integrality the secondary task increases processing demands within the domain of the primary task. Therefore, in the case of dual-task integrality resource reciprocity is not obtained, but instead the resource function in both tasks is identical.

Dual-task integrality has been described on two levels. On a performance level, dual-task integrality results in a facilitation in the performance of one or both tasks when executed concurrently. Facilitation is relative to conditions in which the two tasks are performed separately or when the stimulus relations but not the processing requirements change between dual-task pairs. On a resource level, dual-task integrality occurs when two tasks can be processed within the same resource framework. Thus, there appear to be at least two different types of dual-task combinations that do not result in performance tradeoffs. As argued by capacity theories, tasks which require different processing resources can be sucessfully time shared. In the present study we are suggesting that dual-task decrements can also be avoided if the two tasks permit integrated processing even if the tasks require the same type of processing resources.

Several factors have been proposed to influence the degree of integrality between tasks. One factor, the redundancy between components of the tasks, has been described above. In the context of the present study it is of interest to know whether the correlation between task components has an effect on dual-task processing. Theories of attention have emphasized the influence of the spatial location of stimulus attributes on the efficiency of processing. Treisman (1977), in her Feature Integration Model of Attention, has argued that features which occur within the same central fixation of attention are combined to form a single object. Once the object has been formed it is perceived and stored in memory as such. Kahneman and co-workers (Kahneman & Henik, 1981; Kahneman & Treisman, 1984; Kahneman & Chajczk, 1983; Kahneman, Treisman & Burkell, 1983) have underscored the importance of the object in attention by suggesting that attentional competition arises between, not within object files. This arguement implies that tasks which require the processing of different attributes of the same object will be procesed within the same resource framework. Tasks which necessitate the processing of separate objects will compete for processing resources.

The P300 component of the ERP has been found useful in providing information concerning the allocation of resources to concurrently performed tasks. P300's elicited by discrete secondary task events decrease in amplitude with increases in the difficulty of the primary task (Isreal et al, 1980; Kramer, Wickens & Donchin, 1983). The secondary task methodology assumes that changes in primary task difficulty will be reflected in secondary task performance. Increasing the difficulty of a primary task is presumed to consume resources which would have normally been used in the processing of the secondary task. Thus, the secondary task P300's mirror the proposed resource function. If P300 does in fact reflect the resource structure of dual-tasks then it would be predicted that P300s elicited by discrete primary task events would increase in amplitude with increases in the difficulty of the primary task. This hypothesis was confirmed in a study in which P300s were elicited by discrete spatial changes in the position of a tracking target (Wickens, Kramer, Vanasse & Donchin, 1983). Increasing the difficulty of the tracking task by incrementing the order of the control dynamics resulted in a systematic increase in P300 amplitude. P300s will be employed in the present study to provide information concerning the resource framework of the dual-task combinations.

METHOD

Subjects

Twelve right handed persons (6 male and 6 female) were recruited from the student population at the University of Illinois and paid for their participation in the study. None of the students had any prior experience with the pursuit step tracking task. All of the subjects had normal or corrected to normal vision.

Step Tracking and Discrimination Tasks

The tracking display which consisted of the computer driven target and the subject controlled cursor was presented on a Hewlett Packard CRT which was positioned approximately 70 cm from the subjects. The target and cursor were 1.2 cm x 1.2 cm in size and subtended a visual angle of 1.0 degrees.

The target changed its position along the x axis once every 3.6 to 4 sec and the subjects' task was to nullify the position error between the target and cursor. The cursor was controlled via the manipulation of a joystick with the right hand. Pursuit step tracking was defined as the primary task. The dynamics for the tracking stick were composed of a linear combination of first and second order components. The system output, X(t), is represented by the following equation.

$$X(t) = [(1-a)\int u(t)\,dt] + [(a)\iint u(t)\,dt]$$

where: u = stick position; t = time and a = difficulty level. The task was conducted at three different levels of the system order manipulation: (1) in the relatively easy condition a was set to zero, a pure first order (velocity) system, (2) in the moderate difficulty condition a was set to .5, a 50/50 combination of first and second order dynamics, and, (3) in the difficult tracking condition a was set to 1.0, a pure second order (acceleration) system.

The subjects secondary task involved counting the total number of occurrences of a relevant probe. Probes were presented in a Bernoulli series. The probability of either of the stimuli occuring on any one trial was .50. In different experimental blocks, subjects counted the bright flashes of a horizontal bar, bright flashes of a cursor, translational changes of the cursor upward or translational changes of a horizontal bar downward (see Figures 1 and 2). The two types of stimulus events (brightness and translational changes) were equated for difficulty prior to the experiment. Secondary task probes occurred either on the same x axis as the tracking task or 2 cm (1.5 degrees of visual angle) below it. A probe was presented every 3.6 to 4 sec. The presentation of the probe was temporally constrained so that it occured 1.8 to 2 sec subsequent to a step change in the tracking target. Thus the temporal sequence of the presentation of the probes (secondary task stimuli) and changes in the spatial position of the tracking target was fixed, while the temporal interval between these stimuli was variable.

In the dual-task blocks subjects performed both the tracking and the count tasks. At the conclusion of each block of trials subjects reported their total count. At this time subjects also rated the subjective difficulty of the block on a bipolar scale from 1 (easy) to 7 (difficult). Following each block the subjects were informed of their count accuracy and root mean square tracking error.

Recording System

EEG was recorded from three midline sites (Fz, Cz and Pz) and referred to linked mastoids. Two ground electrodes were positioned on the left side of the forehead. Burden Ag-AgCl electrodes affixed with collodion were used for scalp and mastoid recording. Beckman Bipotential electrodes, affixed with adhesive collars, were placed below and supra-orbitally to the right eye to record electro-oculogram (EOG) and this type of electrode was also used for ground recording. Electrode impedances did not exceed 5 kohms/cm.

The EEG and EOG were amplified with Van Gogh model 50000 amplifiers (time constant 10 sec and upper half amplitude of 35 Hz, 3dB octave roll-off). Both EEG and EOG were sampled for 1280 msec, beginning 100 msec prior to stimulus onset. The data was digitized every 10 msec. ERP's were filtered off-line (-3dB at 6.29 Hz, 0dB at 14.29Hz) prior to statistical analysis. Evaluation of each EOG record for eye movements and blinks was conducted off-line. EOG contamination of EEG traces was compensated for through the use of an eye movement correction procedure (Gratton, Coles &

Donchin, 1982).

Design

A repeated measures, four way factorial design, was employed. The factors were primary task difficulty (count only, first order, first/second order and second order control dynamics), the relationship between primary and secondary task stimulus objects (same or different objects), the spatial position of the primary and secondary tasks (same or different) and the type of secondary task (intensity or translational discriminations). The degree of correlation between the primary and secondary tasks was also manipulated, although this manipulation was not orthogonal to the other four factors. Subjects performed the dual tasks with either 0 or .85 correlation at each level of difficulty in the same object - same position and different object - same position conditions with the intensity discrimination secondary task.

Procedure

Each of the twelve subjects participated in all of the experimental conditions. One practice and two experimental sessions, run on successive days, were required to complete the experiment. The practice session included 24 blocks of tracking and six secondary task count blocks. Each of the tracking blocks lasted four min. Subjects performed eight blocks of tracking at each of the three levels of system order. Secondary task blocks lasted approximately six min.

The experimental sessions began with three single task tracking blocks, each lasting approximately 3 min. Following the single task tracking blocks, subjects performed 15 dual-task blocks. The 30 dual-task blocks divided between sessions 2 and 3 consisted of 24 blocks from the (3 tracking difficulty levels x 2 types of stimuli x 2 task positions x 2 secondary tasks) factorial design and 6 blocks in which dual-tasks in the same object - same position and different object - same position conditions were highly correlated (.85). Each of the dual-task blocks lasted approximately 6 min. Subsequent to the dual-task blocks subjects again performed three single task tracking blocks. ERPs, subjective ratings and RMS tracking error were recorded during the experimental sessions. The order of the experimental blocks was counterbalanced across subjects.

RESULTS

Performance Measures and Subjective Ratings

Figure 3a presents the RMS tracking error for each level of system order during dual-task performance. The figure suggests that increasing system order results in increases in subjects' tracking error. Planned comparisons indicated that subjects performed significantly better with first order than they did with first/second order tracking ($F(1,11)=5.64$, $p<.05$). Performance was also better in the first/second order condition than it was during second order tracking ($F(1,11)=8.58$, $p<.05$). The effect of system order on RMS error did not differ significantly across dual-task, single task or correlated tracking conditions. Thus, the secondary task did not intrude on primary task performance.

Although the secondary task did not affect the RMS error - system order relationship in the single and dual-task tracking blocks, the type of secondary task object did influence the subjects' tracking error. The effect of secondary task object on RMS error is illustrated in Figure 3a. Subjects tracking error was significantly lower when the secondary task involved counting flashes or translational changes of the cursor than when subjects were required to perform the secondary task by counting changes in the horizontal bar (F(1,11)=7.1, p<.05). The differential effect of the type of secondary task object on tracking performance may be due to the relationship of the objects to the primary task. The cursor is clearly a necessary component of the tracking task while the horizontal bar is not necessary for primary task performance. Thus, subjects may find it more difficult to track and count probes if the probes are extraneous to the tracking task than if the probes occur within the primary task. If this interpretation is true we would expect that integration of the two tasks, by correlating events in the primary task with events in the secondary task, would reduce the differences in RMS error between the two objects. A comparison of the correlated and uncorrelated dual-task pairs supports this interpretation. There was no significant difference in RMS error between the horizontal bar and cursor conditions when the primary and secondary tasks were correlated.

Average ratings of difficulty for each level of system order in the dual-task conditions are presented in Figure 3b. Subjects' perception of difficulty increases from the single task count condition to the dual-task conditions as well as with increases in system order within the dual-task conditions (F(3,33)=44.39, p<.001). Subjects rate the difficulty of the dual-tasks higher when performing the secondary task with the horizontal bar than they do when counting the intensity or translational changes of the cursor (F(1,11)=13.84, p<.01). Subjective ratings of difficulty did not differ between objects in the correlated dual-task conditions. Thus, subjects ratings of tracking difficulty are consistent with their overt performance, as measured by RMS tracking error.

The accuracy with which subjects counted the secondary task probes was not significantly affected by any of the experimental manipulations. Subjects' counting accuracy exceeded 97 percent in all of the experimental conditions. Thus, the changes in P300 amplitude as a function of system order cannot be attributed to the subjects failure to accurately count the probes during higher order tracking.

Event-Related Brain Potentials

The treatment of the ERP data is divided into two sections. The first section examines the ERPs elicited by changes in the spatial position of the tracking target. The second section is concerned with the effects of the experimental manipulations on the ERPs elicited by the secondary task probes in the correlated and uncorrelated dual-task conditions.

Primary Task Events Figure 4 presents the ERPs elicited by changes in the spatial position of the tracking target in the dual-task conditions for the parietal recording site. It is evident that the ERPs differ in the amplitude of the positive components as the difficulty of the primary task is varied. This amplitude difference appears as early as 350 msec after the stimulus and continues to the end of the recording epoch. The largest positivity is elicited when tracking is the most difficult.

The ERPs acquired in the dual-task conditions were quantified by averaging the single trials within experimental conditions and analyzing these averages by a Principal Components Analysis (PCA) technique (Donchin & Heffley, 1979). The magnitude of "P300" component increased from Fz to Pz (F(2,22)=115.08, p<.001) and the component loadings were maximal in the epoch associated with P300 (450 - 700 msec). The amplitude of the P300 was influenced by the system order of the tracking task. Increases in system order produced increases in the amplitude of the P300 component (F(2,22)=12.84, p<.001). Thus, consistent with previous research, the amplitude of the P300s elicited by discrete changes in a primary task increase with increases in the difficulty of that task.

Secondary Task Probes: Uncorrelated Dual-Tasks Figure 5 presents the average parietal ERPs elicited by the secondary task probes during the performance of the pursuit step tracking task. Several aspects of the waveforms are noteworthy. In all of the experimental conditions the single task count block elicits a large positivity at approximately 400 msec post-stimulus. This positive deflection has been identified as the P300 component. The three levels of system order elicit varying degrees of positivity which appear to depend on the particular experimental condition. For example, for all experimental conditions in which the cursor is the secondary task probe the waveforms are most positive for the second order condition, of intermediate amplitude in the first/second order condition and smallest in amplitude in the first order condition (F(2,22)=28.1, p<.001). This sequence of levels of system order does not appear to be influenced by the position of the secondary task probe relative to the tracking task or the type of discrimination required of the subject. In the two conditions in which the horizontal bar is located below the tracking task the sequence of the ERPs elicited by different levels of system order is clear and consistent. However, the order is the inverse of that obtained in the cursor conditions. The first order tracking condition elicits the largest amplitude positivity, the first/second order condition elicits an intermediate level of positivity and the second order condition produces the smallest amplitude (F(2,22)=24.2, p<.001). When the horizontal bar is superimposed on the tracking task the ERPs elicited by different levels of system order are not significantly different.

The reciprocial relationship between the cursor and horizontal bar conditions as a function of system order was predicted on the basis of the resource structure inferred from the Object File Model of Attention (Kahneman & Henik, 1981). It was hypothesized that if two tasks required processing of different attributes of the same object then the resource structure of the two tasks would be similar. The direct relationship between P300 amplitude and system order for the primary task events and cursor probes is consistent with this hypothesis. It was also argued that if two tasks required the processing of different objects and these tasks overlapped in their resource demands as defined by the Multiple Resource Model, then the P300 amplitude - system order relationship would be reciporical between primary and secondary tasks. This hypothesis was confirmed with the dual-task combination of the tracking task and horizontal bar. Thus, the results obtained in the present study are consistent with both hypotheses concerning the resource structure of dual-tasks. Tasks which require the processing of different attributes on the same object lead to the same P300 amplitude - system order relationship while two tasks which require the processing of different objects result in a reciprocial P300 amplitude - system order relationship between tasks.

Secondary Task Probes: Correlated Dual-Tasks

Figure 6 presents the average parietal ERPs elicited by the correlated and uncorrelated dual-task conditions. There are several interesting aspects of the waveforms. A comparison of the ERPs elicited in the correlated and uncorrelated cursor probe conditions suggests that system order has the same effect on the ERPs in both conditions. The ERPs elicited by the cursor probes during second order tracking possess a large positive amplitude. The first/second order condition waveforms are of intermediate amplitude and the first order condition ERPs are smallest in amplitude ($F(2,22)=10.9$, $p<.001$). An examination of the waveforms elicited by the horizontal bar probes presents a different picture. The effect of system order is not significant in the uncorrelated horizontal bar condition. However, the ERPs elicited in the correlated horizontal bar condition increase in positivity with increases in system order ($F(2,22)=12.3$, $p<.001$). Thus, it appears that the effect of system order on the ERPs is the same across the two cursor conditions and the correlated horizontal bar condition.

These results suggest that when two tasks are already being processed within the same resource framework, as was the case for the uncorrelated dual-task cursor condition, correlation does not have a large effect on the resources allocated to the tasks. The P300 amplitude - system order relationship was not significantly different in the correlated and uncorrelated dual-task conditions. Thus, when the two tasks require the processing of different attributes on the same object, the processing of the tasks is in some sense integrated and inter-task correlation does not appear to enhance this integrality further. However, when two concurrently performed tasks require the processing of separate objects, as was the case in the horizontal bar conditions, inter-task correlation does appear to enhance the integrality between tasks. This increase in dual-task integrality is inferred from the change in the P300 amplitude - system order relationship between the correlated and uncorrelated horizontal bar conditions. The P300 amplitude - system order relationship in the correlated condition is similar to that obtained for the primary task events suggesting an overlap in the resource structure between tasks.

GENERAL DISCUSSION

In most dual-task combinations increasing the difficulty of one task is assumed to consume resources which normally would be employed in the processing of the other task. The resources shared by these two tasks are presumed to be reciprocal in nature. However, under conditions of dual-task integrality, the secondary task increases processing demands within the domain of the primary task. Therefore, in the case of dual-task integrality, resource reciprocity is not obtained.

The present study represents an initial investigation of some of the factors believed to influence the degree of integrality between tasks. Under conditions of low dual-task integrality, P300s elicited by discrete secondary task events decreased in amplitude with increases in the difficulty of the primary task. The changes in P300 were used to infer changes in the allocation of resources between tasks; increasingly smaller P300s indicating fewer resources available for the secondary task. Thus, when two tasks require similar resources and the processing of the tasks is not integrated, resource reciprocity occurs. In cases in which the processing of two concurrently performed tasks is highly integrated, the P300s elicited by the secondary task events increase in amplitude with increases in the difficulty of the primary task. Thus, the P300s elicited by the secondary task events produce the same amplitude pattern as the primary task P300s.

Three factors were suggested to foster dual-task integrality. It was proposed that tasks which required the processing of different attributes of the same object would result in dual-task integrality while tasks which required the processing of separate objects would result in resource reciprocity. Figure 7 shows that P300s elicited in the same object condition (cursor) increased in amplitude with increases in primary task difficulty. Conversely, P300s elicited by a different secondary task object (horizontal bar - below) decreased in amplitude with increases in the difficulty of the primary task. A second factor proposed to influence the degree of integrality between tasks was inter-task correlation; higher correlation resulting in greater dual-task integrality. P300s elicited by secondary task events which were highly correlated with events in the primary task increased in amplitude with increases in primary task difficulty. Thus, the results confirmed the predictions for both the object and correlation factors. The physical proximity of tasks was also proposed to influence the degree of integrality between tasks. Integrality was predicted to increase with increases in the physical proximity of tasks. This result is confirmed by the P300s elicited in the condition in which the horizontal bars are superimposed on the tracking task. However, the physical proximity of tasks does not have as strong an influence on the degree of integrality between tasks as the other two factors.

Figure 8 presents a model of the processing framework underlying the phenomenon of dual-task integrality as inferred from measures of P300 amplitude. Each of the three stimuli possesses a number of attributes. The subjects are instructed that some of the attributes are task relevant and require processing while other attributes are not neccessary for successful performance of the tasks. The relevant attributes are assigned a high processing priority while other attributes receive a lower priority. Large P300s are elicited by the attributes which are assigned a high priority, small P300s are elicited by the low priority attributes. The stimulus attributes are then aggregated on the basis of task assignments and priorities. The attributes that are necessary for primary task performance receive a higher processing priority than the attributes for the secondary task. However, secondary task attributes which occur on primary task objects are assigned the same processing priority as primary task attributes. Thus, the processing of the secondary task attributes is done within the domain of the primary task. This process represents the phenomenon of dual-task integrality. Secondary task attributes which do not occur on primary task objects are assigned a lower priority. These attributes receive the resources remaining after primary task processing. This process is referred to as resource reciprocity. Resource reciprocity also depends on the overlap between the resources required for primary task performance and those needed for the performance of the secondary task. If the two tasks require different types of processing resources, resource reciprocity will not occur (Wickens, 1980). Inter-task correlation and spatial overlap of the task relevant attributes increase the integrality between tasks by decreasing the distance between the primary and secondary tasks on the integrality continuum. Inter-task correlation is more influential in this respect than physical proximity.

The resource framework inferred from the P300 provides a theoretical account of the effect of several factors on the phenomenon of dual-task integrality. The results also have implications of a more applied nature. The P300 component has been employed as a measure of cognitive workload. P300s elicited by secondary task stimuli decrease in amplitude with increases in primary task difficulty. P300s elicited by discrete primary task events increase in amplitude with increases in the difficulty of the primary task. The resources allocated to tasks have been inferred from changes in P300 amplitude. The results obtained in the present study suggest that the reciprocal relationship between the primary and secondary task depends on the structure of the dual-task. For example, the P300 amplitude - task difficulty relationship changes from the case in which the two tasks require the processing of different attributes on the same object to the situation in which the two tasks necessitate the procesing of different objects. Furthermore, inter-task correlation and the physical proximity of task relevant attributes also have a significant effect on the resource structure of the dual-task pair. These findings suggest that a reliable analysis of the processing demands of a task can only take place within a theoretical framework. The model of dual-task integrality offers one such framework.

References

Alwitt, L.F. (1981). Two neural mechanism related to modes of selective attention. Journal of Experimental Psychology: Human Perception and Performance, 7, 324-332.

Donchin, E. and Heffley, E. (1979). Multivariate analysis of event related potential data: A tutorial. In D. Otto (Ed.), Multidisciplinary Perspectives in Event-Related Potential Research(pp. 555-572). EPA 600/9-77-043, Washington, D.C.: U.S. Government Printing Office.

Gratton, G., Coles, M.G.H. and Donchin, E. (1982). A new method for off-line removal of ocular artifact. Electroencephalography and Clinical Neurophysiology, 55, 468-484.

Isreal, J.B. (1980). Structural interference in dual-task performance: Behavioral and electrophysiological data. Unpublished Ph.D. Dissertation, University of Illinois.

Isreal, J.B., Wickens, C.D., Chesney, G.L. and Donchin, E. (1980). The event-related brain potential as an index of display monitoring workload. Human Factors, 22, 211-224.

James, W. (1890). Principles of Psychology Vol I. New York, Henry Holt and Company.

Kahneman, D. and Henik, A. (1981). Perceptual organization and attention. In M. Kubovy and J.R. Pomerantz (Eds.), Perceptual Organization(pp. 181-209). Hillside, New Jersey: Erlbaum.

Kahneman, D. and Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman, J. Beatty and R. Davies (Eds.), Varieties of Attention.

Kahneman, D. and Chajczyk, D. (1983). Tests of the automaticity of reading: Dilution of stroops effects by color irrelevant stimuli. Journal of Experimental Psychology: Human Perception and Performance, 9, 497-509.

Kahneman, D., Treisman, A. and Burkell, J. (1983). The cost of visual filtering: A new interference effect. Journal of Experimental Psychology: Human Perception and Performance, 9, 510-522.

Kramer, A., Wickens, C.D. and Donchin, E. (1983). An analysis of the processing demands of a complex perceptual-motor task. Human Factors, 25, 597-622.

Navon, D. and Gopher, D. (1979). On the economy of the human processing system. Psychological Review, 86, 214-255.

North, R.A. (1977). Task components and demands as factors in dual-task performance. Aviation Research Laboratory. Report Number ARL-77-2/AFOSE-77-2. University of Illinois at Urbana-Champaign,

Treisman, A. (1977). Focused attention in the perception and retrieval of multidimensional stimuli. Perception and Psychophysics, 22, 1-11.

Trumbo, D., Noble, M. and Swink, J. (1967). Secondary task interference in the performance of tracking tasks. Journal of Experimental Psychology, 73, 232-240.

Wickens, C.D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), Attention and Performance VIII (pp. 239-254). Hillside, New Jersey: Erlbaum.

Wickens, C.D. and Boles, D.B. (1983). The limits of multiple resource theory: The role of task correlation/integration in optimal display formatting. University of Illinois. Engineering Psychology Laboratory Technical Report EPL-83-5/ONR-83-5.

Wickens, C.D. and Kessel. C. (1979). The effects of participatory mode and task workload on the detection of dynamic system failures. IEEE Transactions on Systems, Man and Cybernetics, SMC-9.

Wickens, C., Kramer, A., Vanasse, L. and Donchin, E. (1983). The performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information processing resource. Science, 221, 1080-0182.

This research was supported by the Air Force Office of Scientific Research under contract F49620-79-C-0233 with Dr. Alfred Fregly as technical monitor.

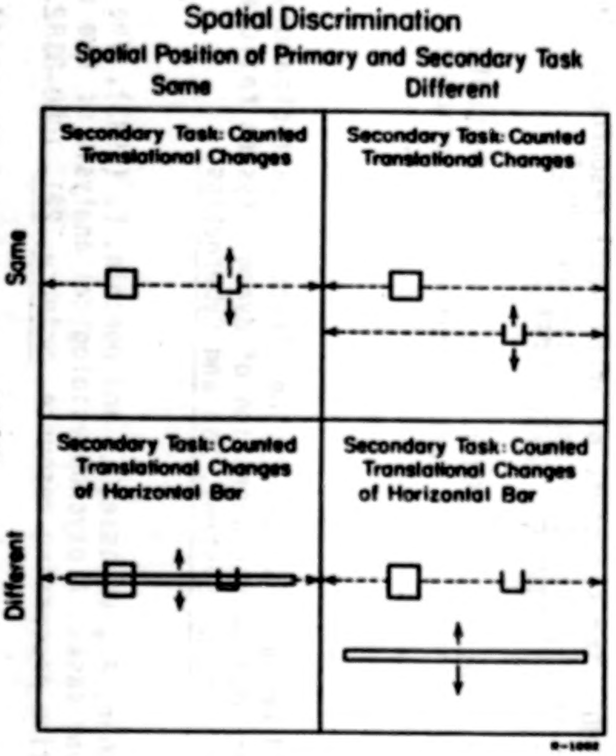

Spatial Discrimination
Spatial Position of Primary and Secondary Task
Same
Different
Objects
Same
Different
Secondary Task: Counted Translational Changes
Secondary Task: Counted Translational Changes
Secondary Task: Counted Translational Changes of Horizontal Bar
Secondary Task: Counted Translational Changes of Horizontal Bar

Figure 1. A graphic illustration of the spatial discrimination secondary tasks and their relationship to experimental manipulations.

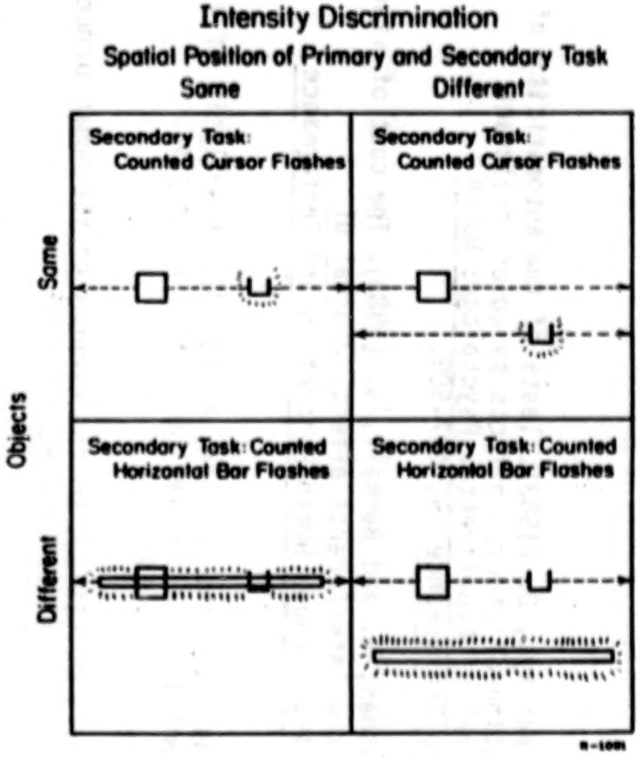

Intensity Discrimination
Spatial Position of Primary and Secondary Task
Same
Different
Objects
Same
Different
Secondary Task: Counted Cursor Flashes
Secondary Task: Counted Cursor Flashes
Secondary Task: Counted Horizontal Bar Flashes
Secondary Task: Counted Horizontal Bar Flashes

Figure 2. A graphic illustration of the intensity discrimination secondary tasks and their relationship to experimental manipulations.

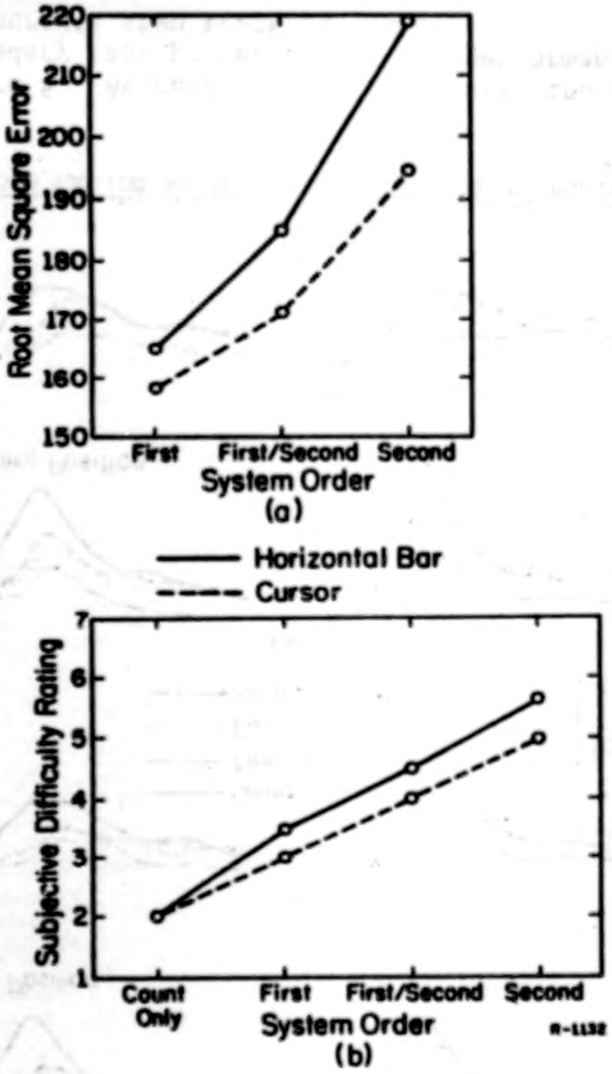


Figure 3. RMS error (a) and subjective difficulty ratings (b) for each level of system order during dual-task performance.

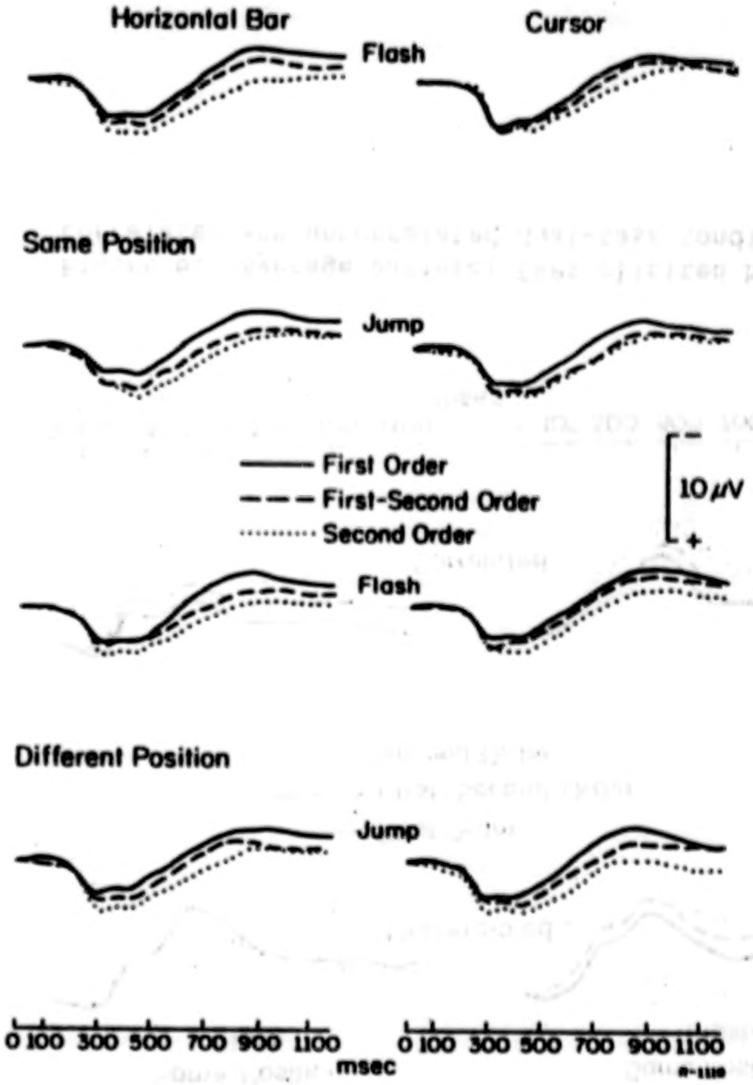


Figure 4. Average parietal ERPs elicited by changes in the spatial position of the tracking target in the dual-task blocks.

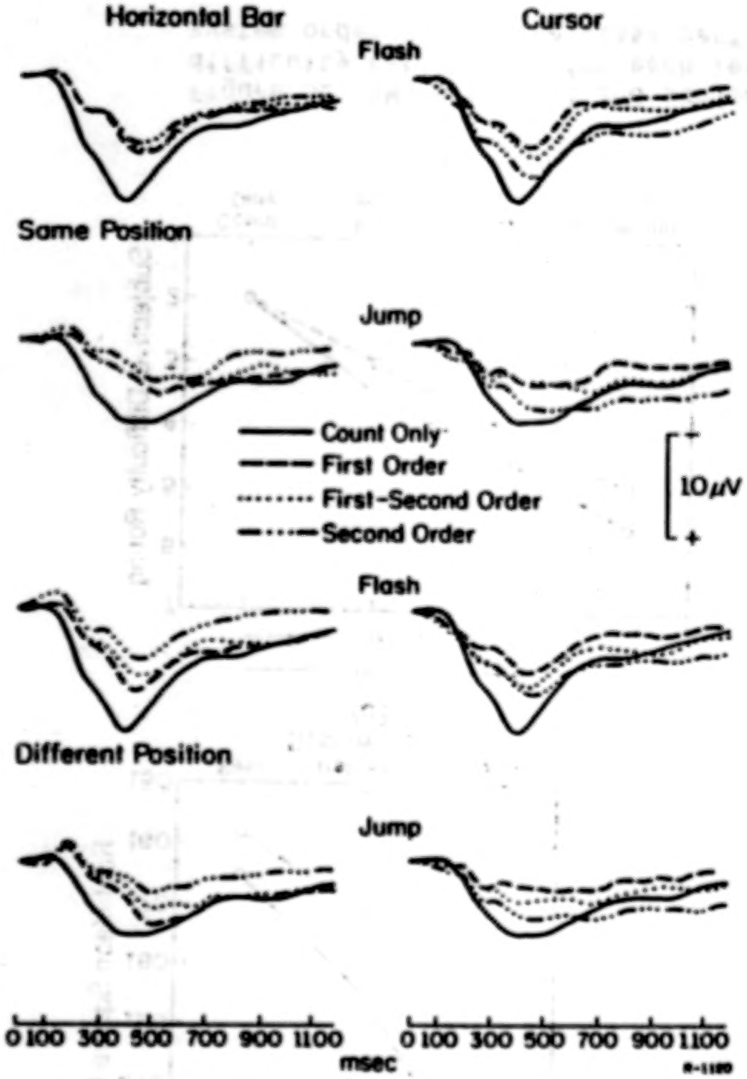


Figure 5. Average parietal ERPs elicited by the secondary task probes during the performance of the pursuit step tracking task.

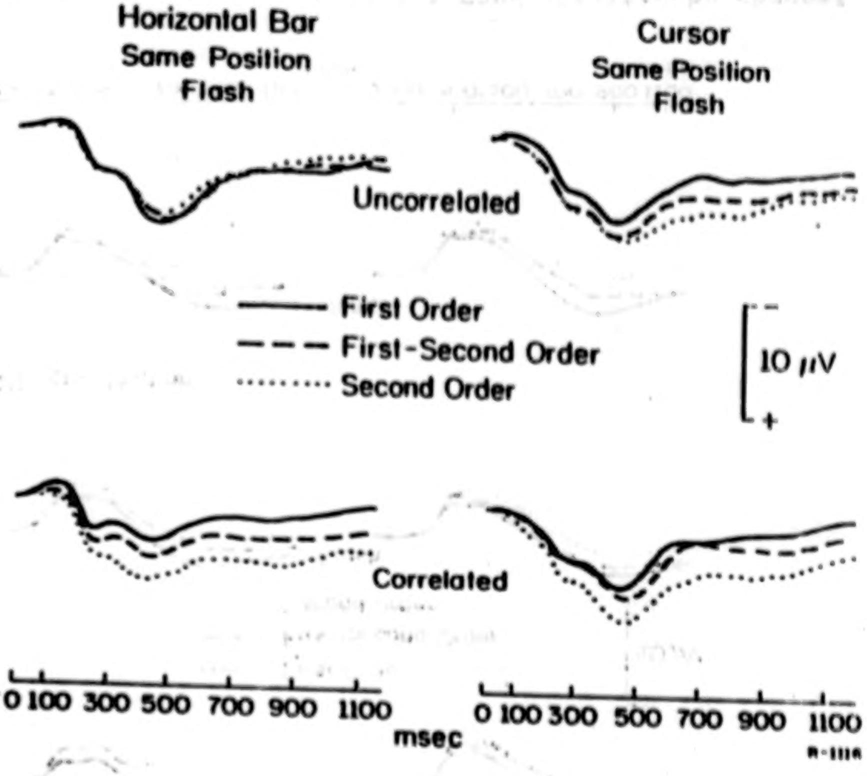


Figure 6. Average parietal ERPs elicited by the correlated and uncorrelated dual-task conditions.

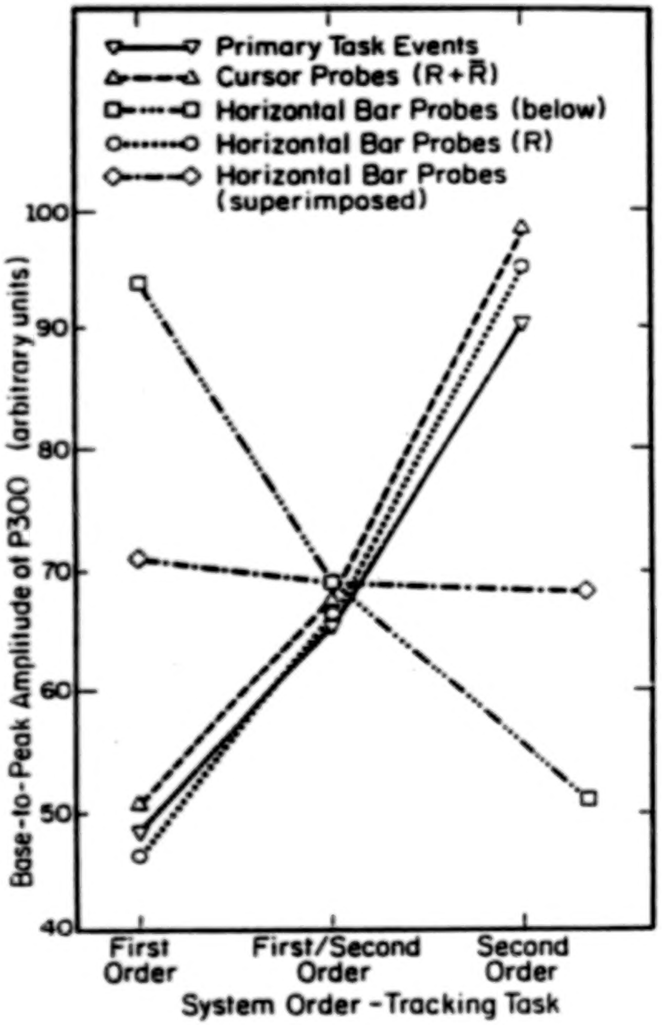


Figure 7. A graphic summary of the P300 results.

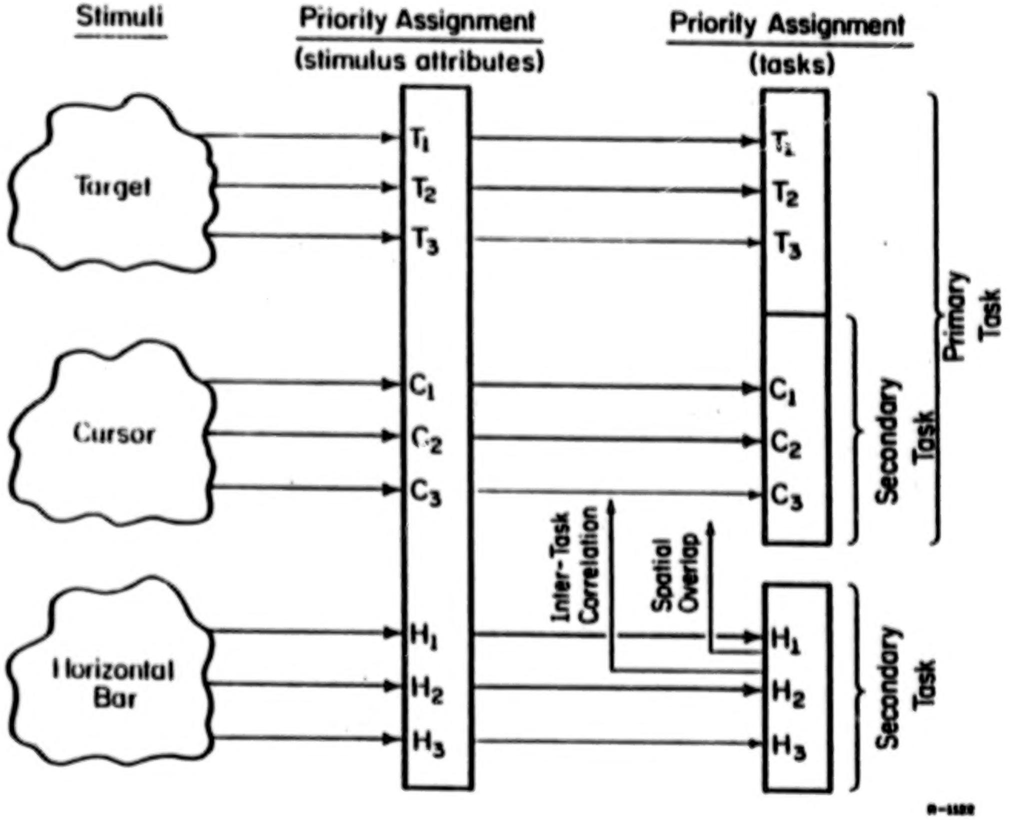

Stimuli
Priority Assignment
(stimulus attributes)
Priority Assignment
(tasks)
Target
Cursor
Horizontal Bar
T1
T2
T3
C1
C2
C3
H1
H2
H3
Inter-Task Correlation
Spatial Overlap
Primary Task
Secondary Task
Secondary Task
R-1122

Figure 8. A model of dual-task integrality inferred from changes in the amplitude of the P300 as a function of primary task difficulty. The subscripted letters represent stimulus attributes. The shade of the attribute lines represent the amount of processing.

N85
14538

UNCLAS

D3

N85 14538

IN SEARCH OF A VISUAL-CORTICAL DESCRIBING FUNCTION
A SUMMARY OF WORK IN PROGRESS

ANDREW M. JUNKER
AIR FORCE AEROSPACE MEDICAL RESEARCH LABORATORY
WRIGHT-PATTERSON AIR FORCE BASE, DAYTON, OHIO 45433

KAREN J. PEIO
SYSTEMS RESEARCH LABORATORIES, INC.
DAYTON, OHIO 45440

INTRODUCTION

By using appropriate signal averaging techniques, it is possible to detect a response in the human EEG to evoking stimuli such as light or sound. When a light stimulus is presented with the light intensity sinusoidally modulated, the result is called a steady state evoked response (SSER). Work done in this area (Regan, 1975; Spekreijse, 1966; Wilson and O'Donnell, 1981; Wilson, 1979) suggests that the SSER may be a useful indicator of internal cortical functioning. Previous work concentrated on the use of a single sine wave or at most three sine waves to drive the evoking stimulus. The thrust of the present work is to explore the utility of using a sum of sinusoids (seven or more) to obtain an evoked response and, furthermore, to see if the response is sensitive to changes in cognitive processing. Within the field of automatic control system technology, a mathematical input/output relationship for a sinusoidally stimulated nonlinear system is defined as a describing function (Kochenburger, 1950). Applying this technology, we have designed our sum of sines inputs to yield describing functions for the visual-cortical response. What follows is a description of the method used to obtain visual-cortical describing functions. A number of measurement system redesigns were necessary to achieve the desired frequency resolution. Results that guided and came out of the redesigns are presented. Preliminary results of stimulus parameter effects (average intensity and depth of modulation) are also shown.

METHOD

Apparatus

A device was constructed which could simultaneously evoke a visual cortical response using flickering lights and provide video driven cognitively demanding tasks (Figure 1). This is accomplished by combining the two images through an 18 cm x 26 cm half-silvered mirror placed at 45 degrees to both images.

PRECEDING PAGE BLANK NOT FILMED

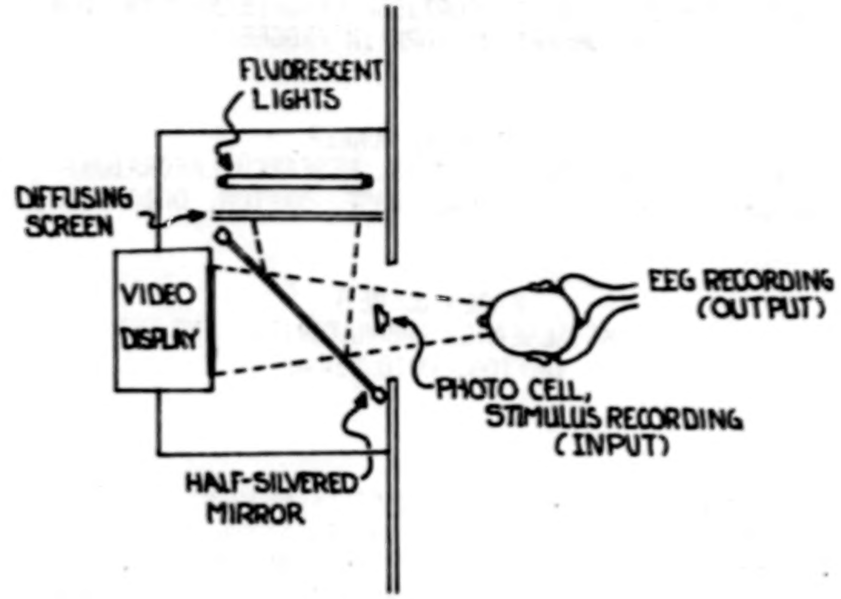


Figure 1. Experimental Setup

The evoking stimulus is provided by two horizontal fluorescent tubes 26 cm long and mounted 4 cm behind a 25 x 27 cm translucent screen in order to distribute light as evenly as possible across the visual field. The average intensity of these lights can be varied from 0 to 160 ft-L. This average intensity range is considered low enough to comfortably present a video display task in the same visual field. A United Detector high speed photo cell is placed 13 cm in front of this display to record amplitude of the input stimuli for comparisons with EEG response. In addition to the image provided by the evoking stimulus, the second image (the video task display) is presented on an Audiometrix 11-inch video monitor.

Stimulus

Sinusoidally modulated lights served as the evoking stimuli. Sums of 10,11 and 13 sine waves were used to modulate these lights around average intensities of 40 and 80 ft-L. Each sine wave provided anywhere from 6.5% to 13% modulation depending on the stimulus parameter chosen. The sine waves were chosen to avoid presentations at frequencies which are harmonics of the other sine waves in the sum. In addition, no one frequency presentation contains a sum or difference of any other sine wave input. These restrictions on sine wave selection are due to the nonlinear behavior of our flickering light generator. Appropriate input selection insures that the nonlinear light effects will not occur where we place the input sine waves. This will also facilitate future investigations of first order nonlinear properties of the evoked response system at harmonics and intermodulation frequencies of the input frequencies (Victor and Shapley, 1980). Our input sine waves range between 5.5 Hz and 49 Hz. Our sine waves are also selected so that they all are multiples of the fundamental frequency (.25 Hz for the first results, .125 Hz for the first redesign, and .0244 Hz for the final redesigned system).

For our preliminary efforts the sum of sines (SOS) was generated by a PDP 11/60 computer and stored on an Ampex SP300 analog tape. Signals from the SP300 were used as input to drive a Scientific Prototype tachistoscope, model G.B., modified so that lamp intensity could be modulated from an external oscillator. For the most recent system design, the PDP 11/60 generated SOS directly drove the tachistoscope.

In addition to the evoking stimuli, a supervisory control task was presented as a cognitive load (Pattipati, Ephraph, and Kleinman, 1975). A manually controlled subcritical tracking task was also used as a possible cognitive driver (Zacharias and Levison, 1979).

Analysis

The describing function is a complex measure of the input-output relationship of a system. We chose to look at this measure in terms of amplitude ratio and phase angle. For our preliminary work a Nicolet Fast Fourier Transform (FFT) analyzer provided these measures. For our final system design we used a PDP 11/60 to both generate the SOS and collect response data, and a PDP 11/34 to perform FFT's. We also computed remnant or background EEG power spectrums. We compute background EEG power by finding the average value of the EEG power within a remnant window (20 frequency bins, 0.488 Hz for the most recent system design) centered about each input frequency of the SOS. Of course the average excludes the power at the center frequency as this is considered the evoked response. For a detailed discussion of guidelines for analysis of frequency response data, see Levison (1983).

Procedure

Subjects were seated in a darkened IAC chamber in front of a 15 cm^2 window and looked into the stimulus presentation device. For the lights only condition, the subjects were instructed to "relax and fixate on a small dot on the center of the display" while the lights were flickering. For the cognitive loading conditions, the subjects were told to concentrate on the tasks.

Recording

Measurements of the evoked response were recorded using silver/silver chloride electrodes at Oz according to the 10-20 international System, with mastoid reference and ground. Resistance between electrodes was less than 5 K ohms. EEG signals from subjects were amplified by Grass P511 AC amplifiers with an effective bandpass of 0.2 to 300 Hz. Sixty Hz filters were not used. Analysis of stimulus and EEG data averaged over 32 4-second epochs (using a 50 percent redundant sliding window) was accomplished with a Nicolet 660 A dual-channel FFT for preliminary results. For the first redesign 16 8-second epochs were used. For the final redesign a PDP 11/34 is used for data analysis. Fourier transforms of 2048 point 40.96 sec time histories are performed.

RESULTS AND DISCUSSION

Preliminary Results

Figure 2 is an illustration of typical power spectrums. The top graph is the spectrum for an evoking stimulus at 10 sine waves as measured at the photo cell. The lower trace is the human EEG response for a lights only condition. Clearly significant peak responses to the input stimuli are indicated in addition to the background EEG.

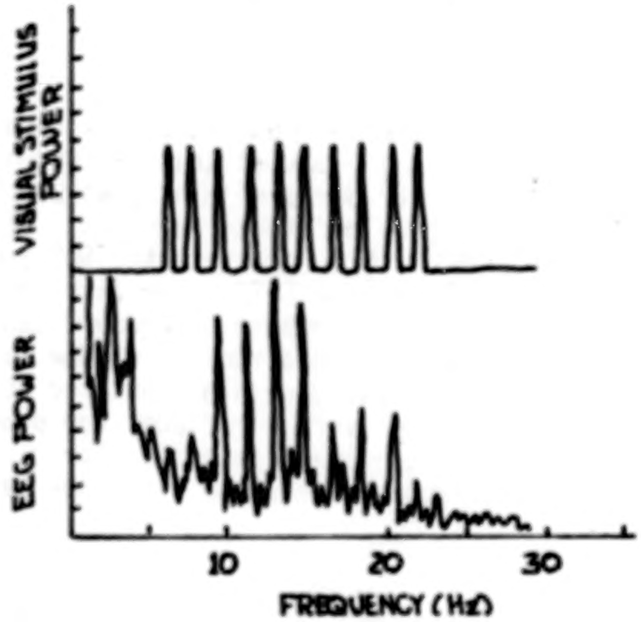


Figure 2. Stimulus and EEG Power Spectrums.

The describing function for a typical subject, which consists of magnitude ratio and phase, is plotted in Fig. 3. The plotted values are for conditions

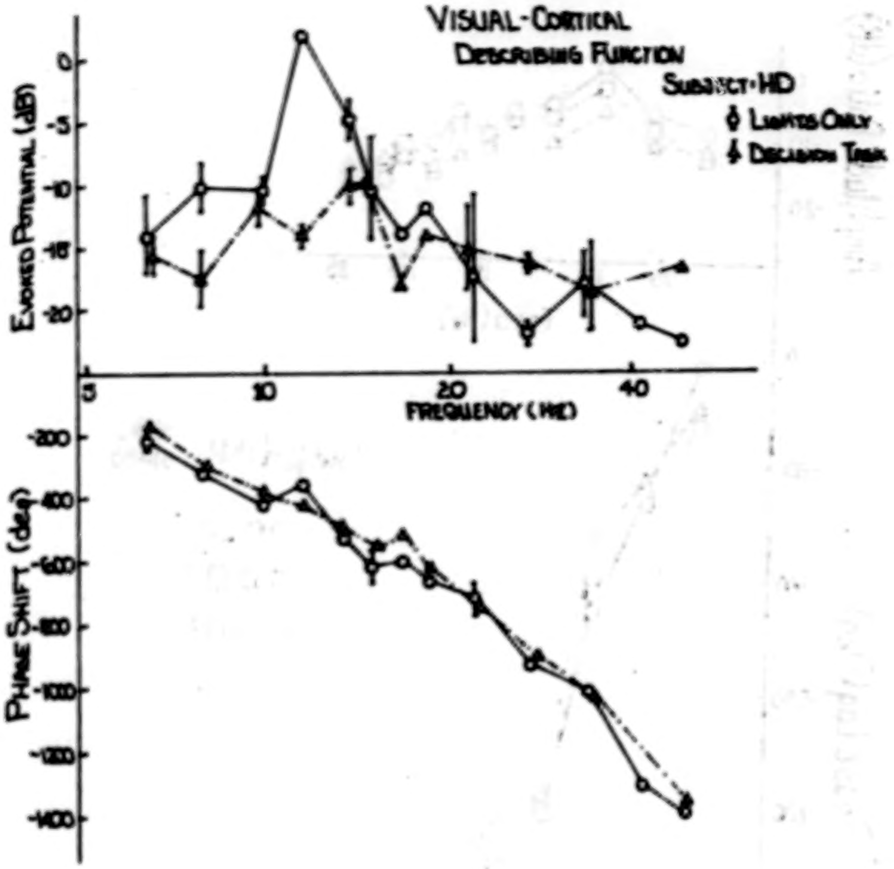


Figure 3. Visual-Cortical Describing Functions.

of lights only and decision-making. They represent the average (indicating mean and standard deviation) of three 68-second runs. Clear differences exist between conditions for frequencies below 22 Hz. A reduction in channel gain is intuitively expected in going from lights only to decision-making. Among other differences between the two phase curves, a decrease in phase steepness for decision-making at 11.5 Hz can be observed. The describing function changes across conditions suggest the notion that the visual-cortical channel changes its dynamic response for different cognitive loads. As a result of these preliminary findings we decided to concentrate our exploration below 25 Hz. In this way we were able to increase our frequency resolution from 0.25 Hz to 0.125 hz. We also reduced the frequency span of our SOS stimulus in hopes of more accurately capturing more subtle gain and phase changes.

Redesign 1

The results of this system redesign are shown for two other subjects in Figures 4 and 5. Evoked response describing functions are compared across

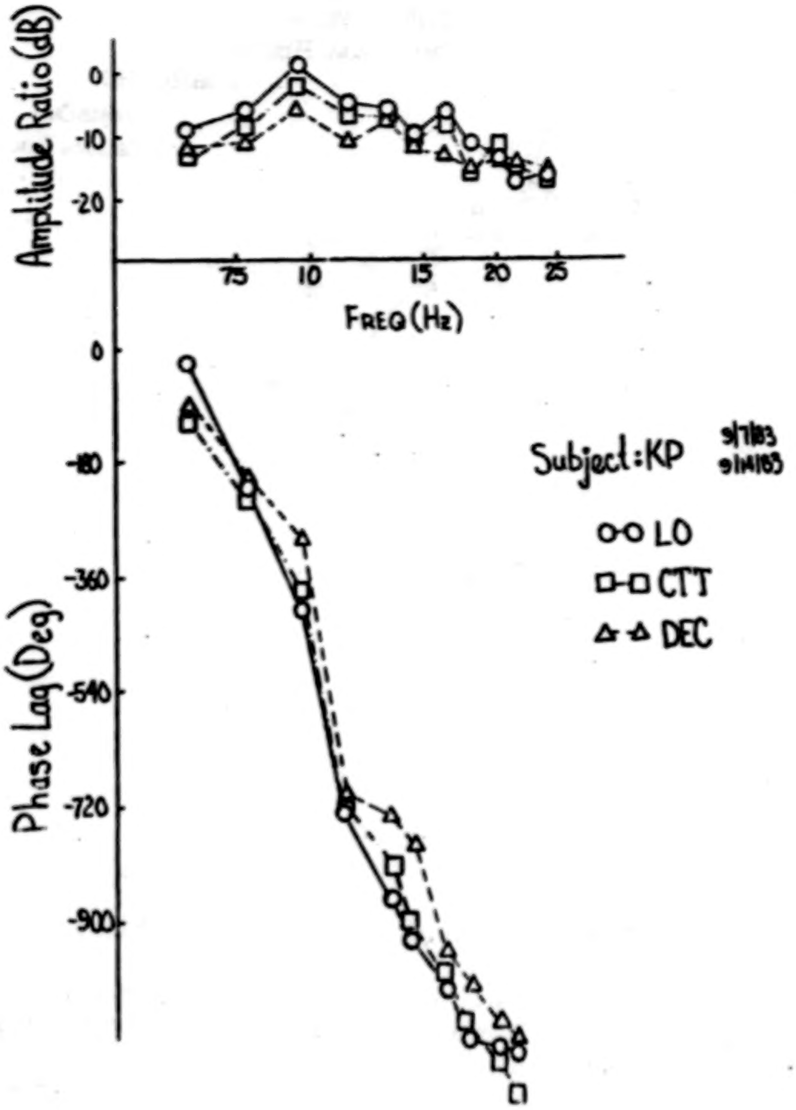


Figure 4. Visual-Cortical Describing Functions.

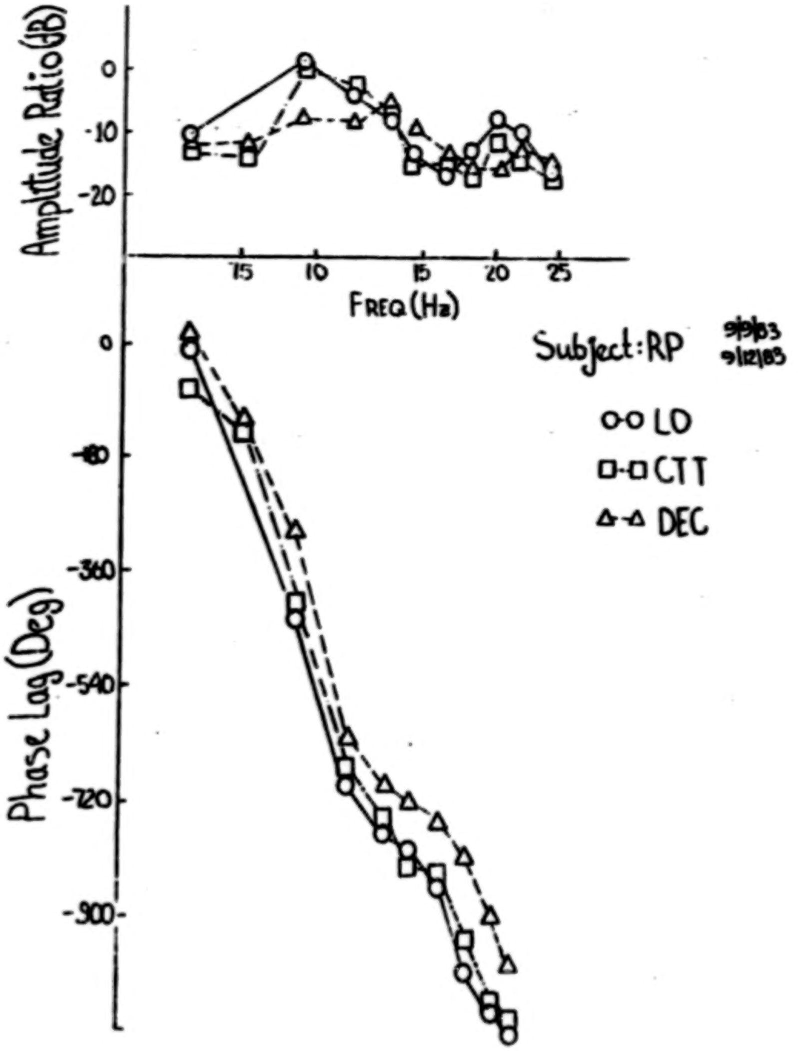


Figure 5. Visual-Cortical Describing Functions.

three conditions; lights only (LO), subcritical manual tracking (CTT) and decision making (DEC). Again similar gain and phase differences between decision-making and lights only exist. Differences between lights only and manual tracking are minimal. This seems reasonable as the tracking task reduces to principally a minimal cognitive load once it is learned.

It is not shown in these graphs, however, the variability of the gain and phase measurements was too great to allow useful modeling of this data. Thus the need for another system redesign was indicated.

Final Redesign

Referring back to figure 2, the input SOS peaks were smeared as they spanned 5 to 7 frequency bins. This smearing was partially due to the Nicolet FFT analyzer's resolution. Furthermore, system syncronization between SOS generation and data collection was lacking. It was hypothesized that these two deficiencies significantly contributed to the extreme variability of the experimental results. To overcome these problems, in the redesign, the computer that generated the SOS (a PDP 11/60) was also used to collect the responses and a PDP 11/34 was used to analyze the collected data. This redesign resulted in a 10 fold increase in frequency resolution over the original system configuration.

The effects of this redesign can be seen in expanded views of input and output power spectrums. The results are plotted for two sine waves; one (18.26 Hz) in figure 6, the other (9.5 Hz) in figure 7. Referring first

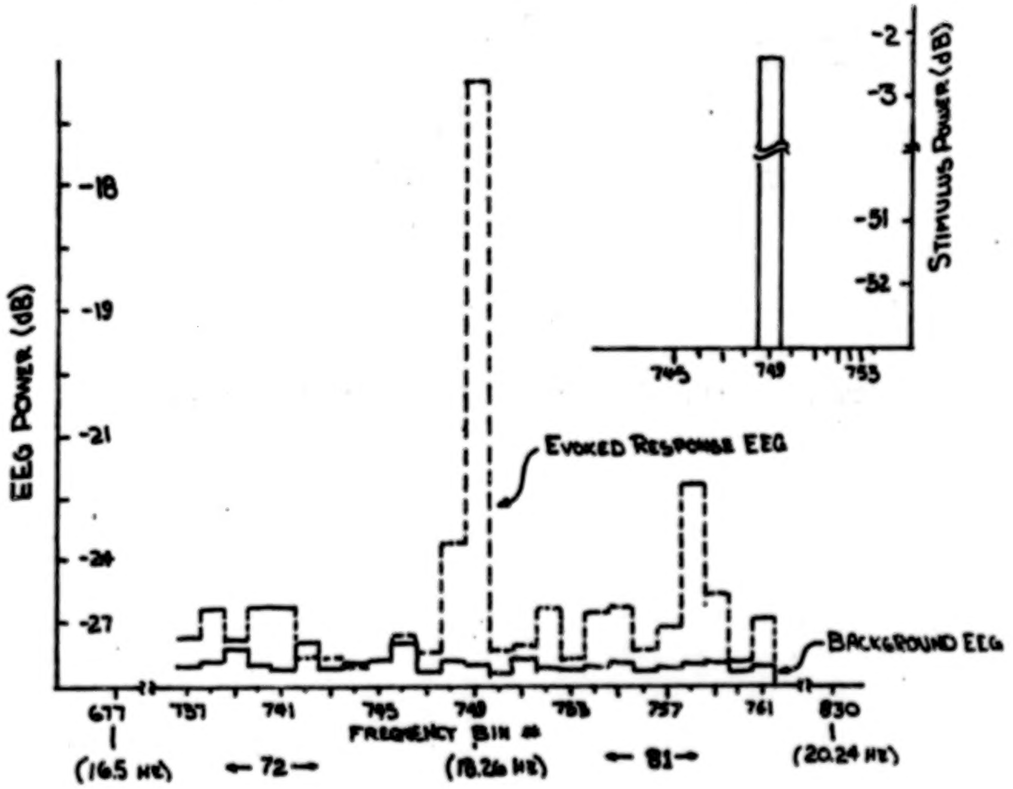


Figure 6. Specificity of Evoked Response (Mid-Frequency Range).

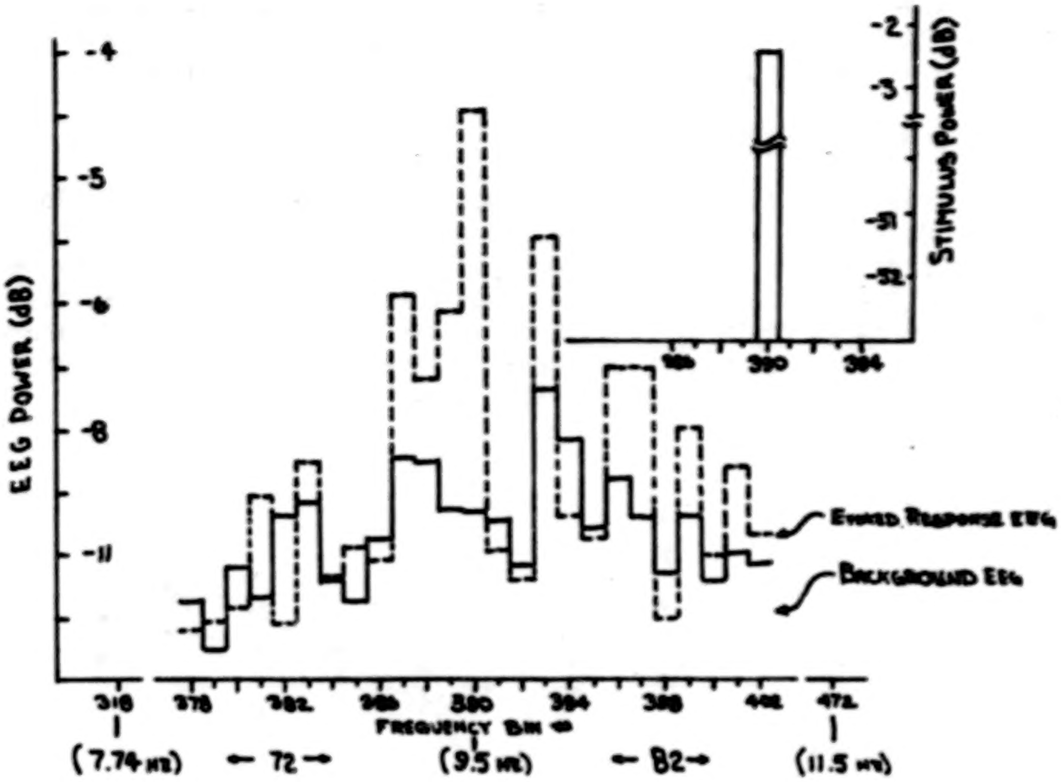


Figure 7. Specificity of Evoked Response (Alpha Freq. Range).

to figure 6, the input (stimulus power) as depicted in the upper right hand graph indicates the excellent stimulus purity achieved. The power appears in one frequency bin of 0.0244 Hz width with adjacent power down 50 dB. Likewise the specificity of the evoked response power is obvious when it is compared to background EEG power levels. To insure that the evoked response was not due to stimulus contamination of the EEG, similar measurements were obtained with the subject blind folded. As expected, there was no response to the stimulus.

Figure 7 shows the effects of the evoking stimulus at 9.5 Hz. This is within the alpha region of the EEG spectrum (8 to 12 Hz). This subject consistently exhibits greatest power within this region. Describing functions and remnant spectrums are plotted for this subject (subject 5) in figures 9 and 10. Because this subject is a large "alpha producer" (greatest evoked response and background EEG levels in the 8 to 12 Hz region) we were interested in seeing what the specificity of the evoked response would be to the stimulus in this region. Evoked response to the stimulus is still significant however there is a significant increase in background EEG power as well. These results indicate that the evoked response and background EEG measurements may be more coupled together in this region (8-12 Hz) than in the mid-frequency region (14-20 Hz).

As part of the final redesign we used the results as shown in figures 6 and 7 to determine the width of the remnant computation window. It was chosen to span 10 frequency bins below and 10 bins above each evoking stimulus frequency. This is the range (0.488 Hz) over which an average EEG power value is computed to yield remnant or background EEG power.

Stimulus Parameter Investigation

With this final system redesign completed we were ready to explore the effects of different stimulus attributes on the evoked response. The two major parameter values that can be manipulated are average intensity level and depth of modulation. We chose 40 and 80 ft-L as average intensity levels, and 13% and 6.5% as depth of modulation values for each of the 10 sine waves in the SOS input. The results of these different parameter values are shown for 4 subjects in figures 8 through 13.

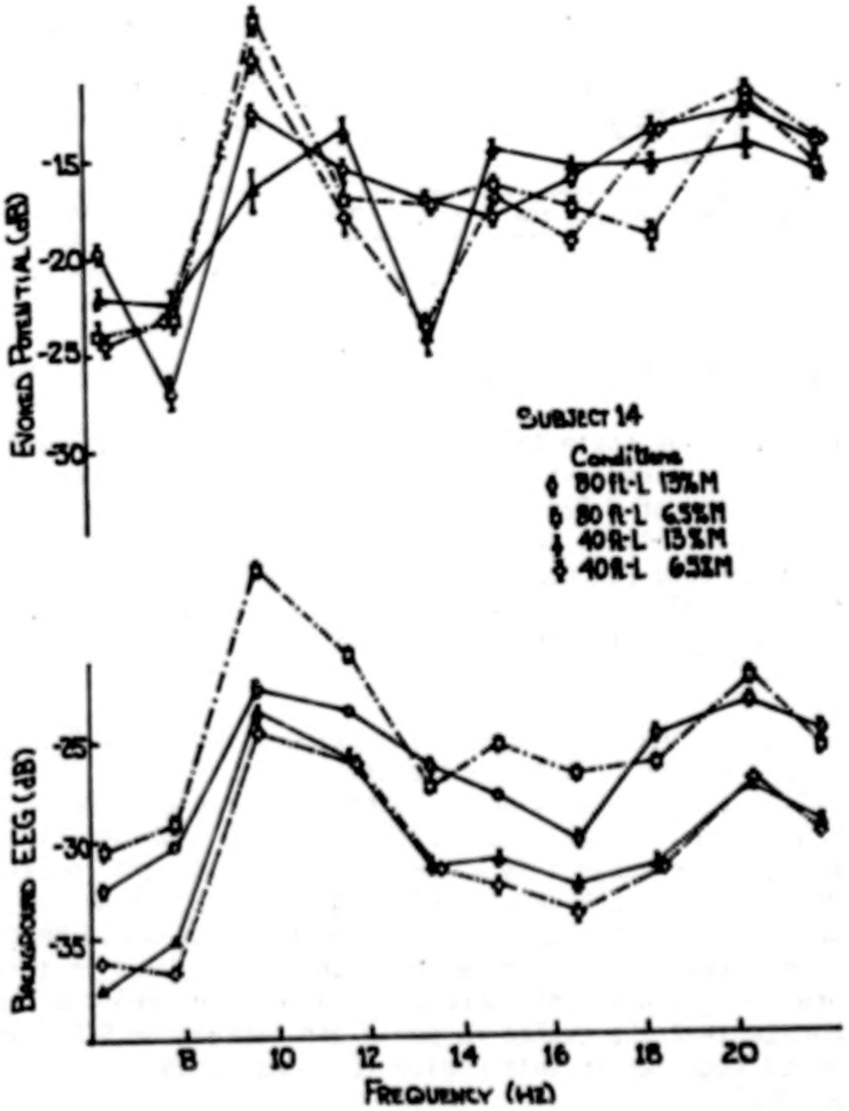


Figure 8. Stimulus Parameter Effects, Subject 14.

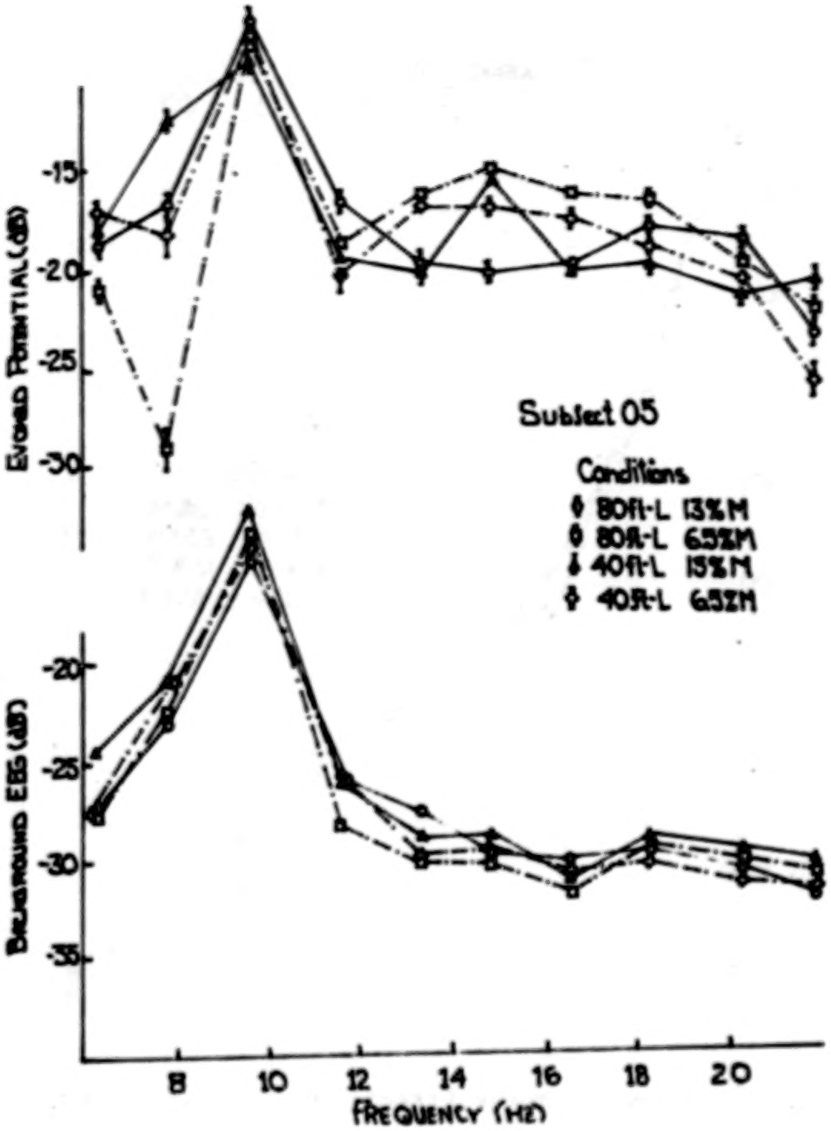


Figure 9. Stimulus Parameter Effects, Subject 5

Each background EEG measurement is the average power taken over 20 bins (10 above and 10 below) centered about each of the 10 frequency values of the SOS stimulus, excluding the power at the centered value, converted to dB. Each evoked potential measurement was computed as follows; average gain was computed from average real and average imaginary FFT components at the stimulus input frequency values, evoked potential was then computed as gain times input and converted to dB. These operations were performed to permit direct comparisons between evoked response and background EEG and across stimulus conditions.

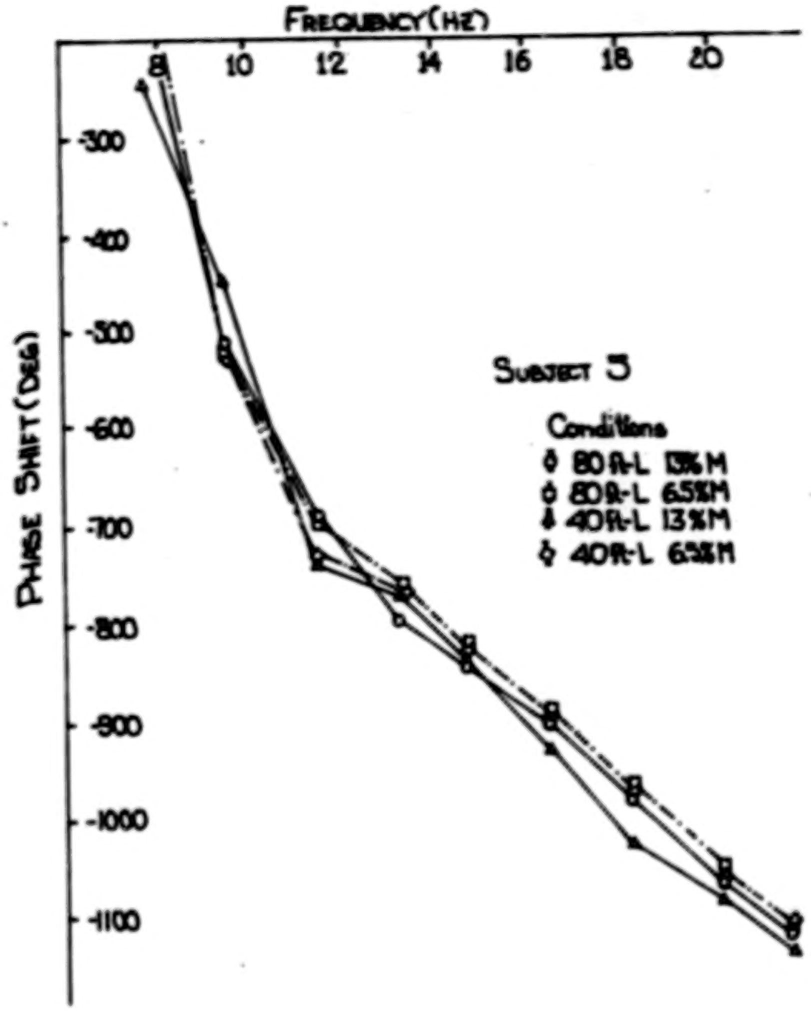


Figure 10. Stimulus Parameter Effects, Phase Shift, Subject 5.

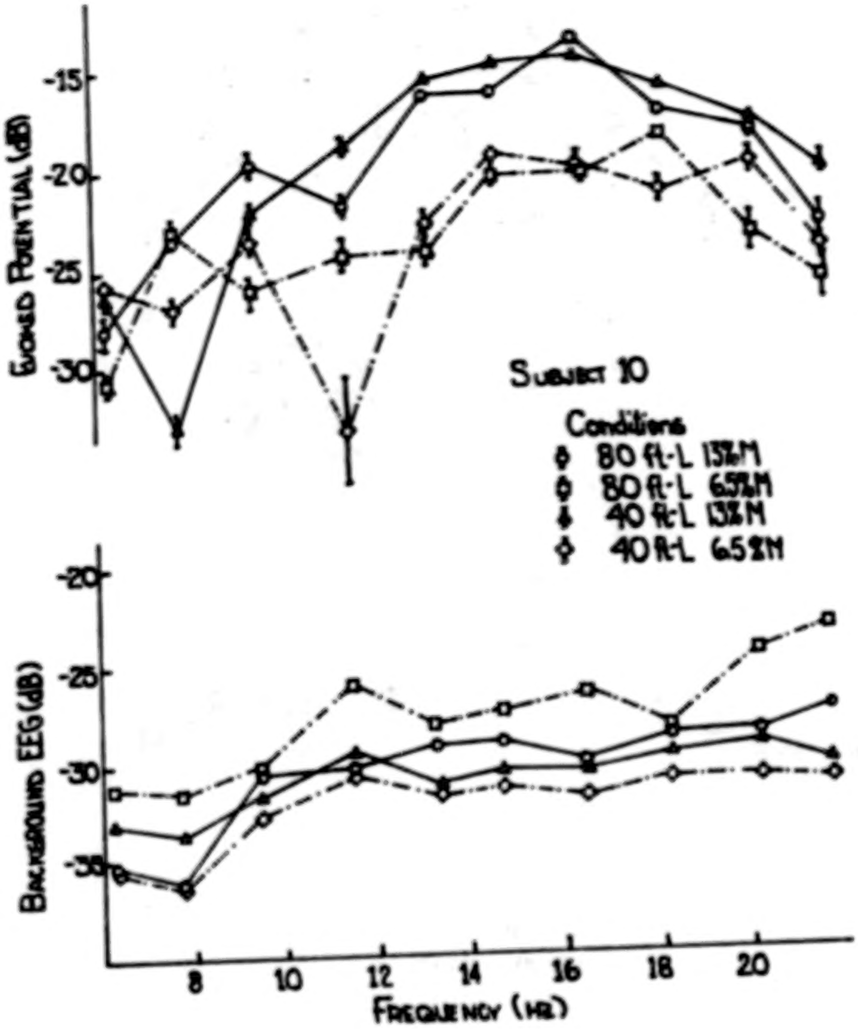


Figure 11. Stimulus Parameter Effects, Subject 10.

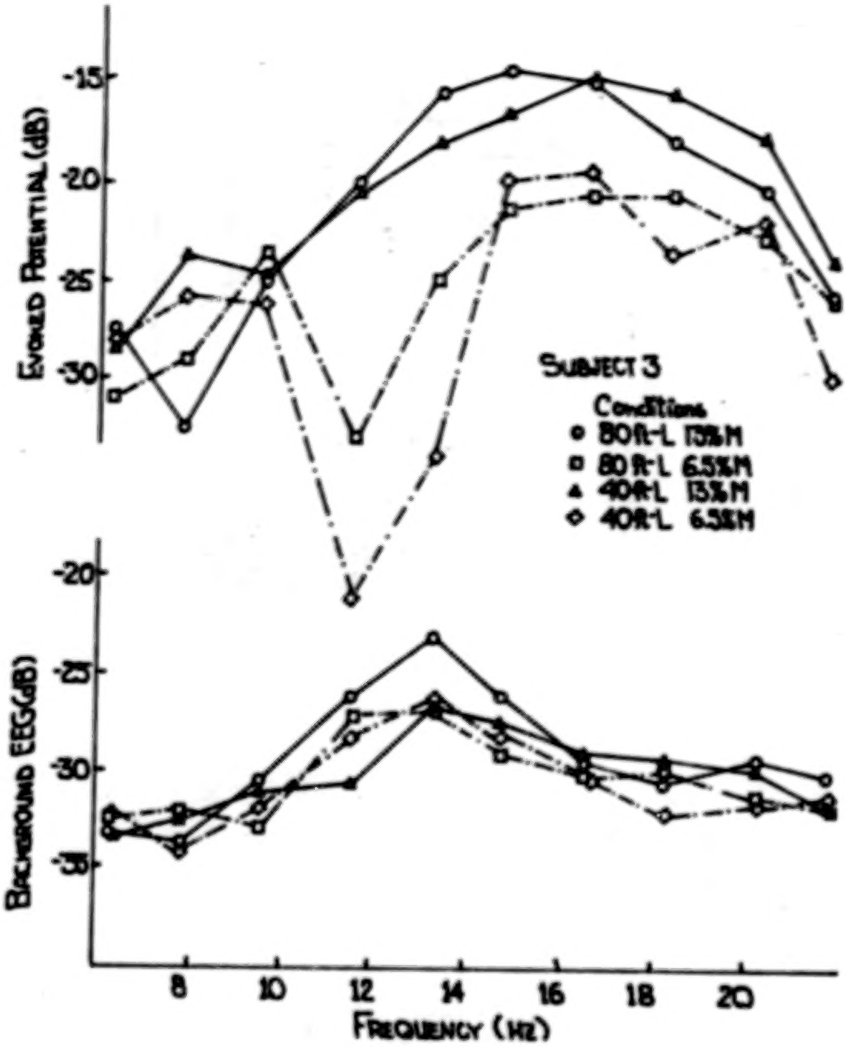


Figure 12. Stimulus Parameter Effects, Subject 3.

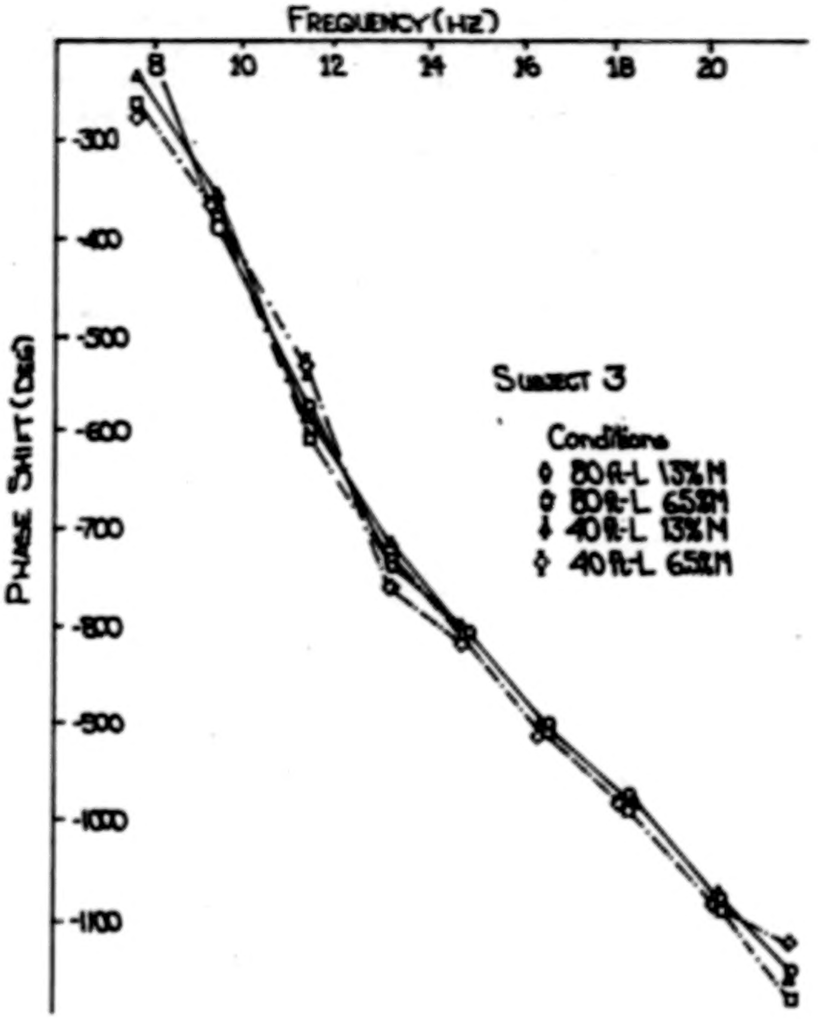


Figure 13. Stimulus Parameter Effects, Phase Shift, Subject 3.

Things to note are the alpha region peaks for subjects 14 and 5 (figures 8 and 9) and the marked absence of alpha region peaks for subject 10 and subject 3 (figures 11 and 12). Note also that different average intensities do not cause consistent evoked response differences. This was expected (Regan, 1975; Spekreijse, 1966). Depth of modulation appears to have an effect on subject 10 and subject 3 (figures 11 and 12). This trend was expected for all subjects however (Regan, 1975; spekreijse, 1966). The most obvious differences between these 4 subjects are the presence or absence of large evoked potential values in the alpha (8 to 12 Hz) region.

To better compare differences between subjects; for each subject we averaged evoked response measures across stimulus conditions, and plotted the averages for each of six subjects in figures 14 and 15. Some things to

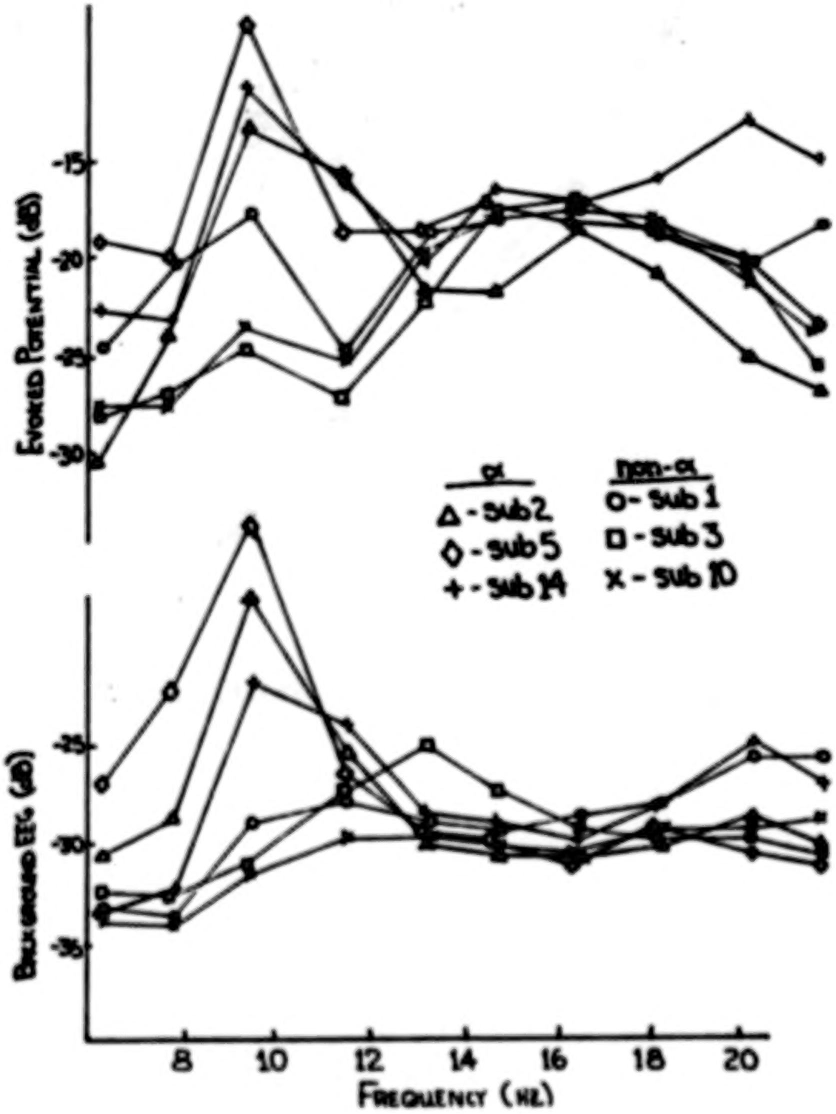


Figure 14. Across Subject Differences.

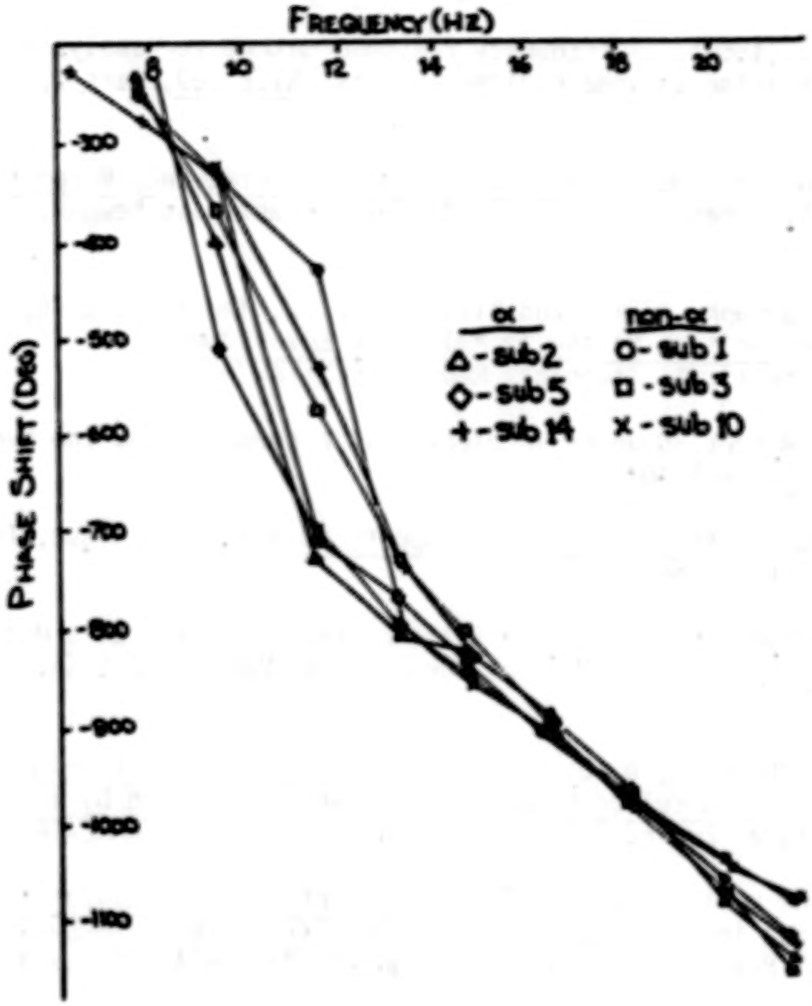


Figure 15. Across Subject Differences, Phase Shift.

note are that across subject differences are least in the mid-frequency region. Three subjects exhibited large alpha region peaks, three did not. All subjects exhibited some alpha region peaking. Phase lags across conditions do not show trends consistent with alpha peaking.

Summary

Results to date continue to support the notion that the SSER is mediated by cognitive processes. The most recent results of the stimulus parameter investigation suggest that the most useful region to explore for cognitive effects may be in the mid-frequency region. Future work will involve modeling evoked response data, determining whether differences found thus far can be explained in a parsimonious manner and whether these differences will be found over more subtle variations in task difficulty or increased cognitive loads.

REFERENCES

Kochenburger, R. J. (1950). A frequency response method for analyzing and synthesizing contractor servomechanisms. Trans. AIEE, 69, Part I, 270-284.

Levison, W. H. (February 1983). Some Guidelines for Frequency Response Analysis. Technical Memorandum CSD 83-4, Bolt Beranak and Newman, Inc.

Pattipati, K. R., Ephraph, A. R., and Kleinman, D. L. (November 1975). Analysis of Human Decision-Making in Multitask Environments. University of Connecticut, Technical Report EECS-TR-79-15.

Regan, D. (1975). Recent advances in electrical recording from the human brain. Nature, 253, 401-407.

Spekreijse, H. (1966). Analysis of EEG Response in Man. The Hague, The Netherlands, Junk Publishers.

Wilson, G. F. (November 1979(. Steady State Evoked Responses as a Measure of Tracking Difficulty. Final report funded by AFOSR Contract No. F49620-79-C.

Wilson, G. F. and O'Donnell, R. D. (1981). Human Sensitivity to High Frequency Sine Wave and Pulsed Light Stimulation as Measured by the Steady State Cortical Evoked Response. Technical Report AFAMRL-TR-80-133.

Zacharias, G. L. and Levison, W. H. (1979). A Performance Analyzer for Identifying Changes in Human Operator Tracking Strategies. AMRL-TR-79-17, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, March 1979.

Victor, J. D. and Shapky, R. M. (1980). A Method of Nonlinear Analysis in the Frequency Domain. Biophys. J. 29, 459-484.

Subjective Evaluation of Workload

N85
14539

UNCLAS

D4

N85 14539

Measurement Of Workload: Physics, Psychophysics, and Metaphysics.

Daniel Gopher
Technion - ITT
Haifa, Israel

The measurement of operator workload is an issue of great concern in the design and evaluation of modern engineering systems. This concern has led to the development of a wide arsenal of measurement techniques, all intended to quantify the phenomena accompanying the behavior of the human processing system when its capacity to meet task demands has been exceeded. Three general categories of measurement approaches are, performance based measures, physiological indices, and subjective scales. In theory, the three approaches should constitute alternative strategies to expose the hidden limitations of internal processors. In practice, there is only a sparse knolledge on the relationship between workload measures obtained under different approaches. Moreover, there appears to be a debate among proponents of these approaches on the validity, comprehensiveness and exclusiveness of different measures.

The present paper reviews the results of two experiments in which workload analysis was conducted based upon performance measures, brain evoked potentials and magnitude estimations of subjective load. The three types of measures were jointly applied to the description of the behavior of subjects in a wide battery of experimental tasks. Data analysis shows both instances of association and dissociation between types of measures. A general conceptual framework and methodological guidelines are proposed to account for these findings.

N85
14540

UNCLAS

N85 14540 D5

SUBJECTIVE WORKLOAD ASSESSMENT AND VOLUNTARY CONTROL OF EFFORT IN A TRACKING TASK

Michael A. Vidulich*
Department of Psychology
University of Illinois at Urbana-Champaign
Champaign, Illinois

Christopher D. Wickens
Institute of Aviation and Department of Psychology
University of Illinois at Urbana-Champaign
Willard Airport
Savoy, Illinois

ABSTRACT

A manual control tracking task was manipulated along two dimensions: (1) control order, and (2) forcing function bandwidth. In the first phase of the experiment subjective workload assessments were collected. It was found that subjective assessments of workload were closely associated with performance in the case of increasing control order, but not in the case of increasing bandwidth. This was interpreted as indicating that subjective workload assessments are most appropriate for the study of increasing difficulty centered in response selection processes as opposed to response execution processes. In the second phase of the experiment the subjects were asked to voluntarily limit the effort they applied in the performance of the tracking task. The results indicate that the subjects were quite facile in doing this. However, comparison of this data to the findings of other studies that manipulated effort via dual-task biasing indicate that effort manipulation is much more potent in a single-task configuration. This finding is discussed in terms of multiple resource theories of attentional capacity. Also, the utility of an analysis of covariance (ANACOVA) procedure in studying the relationships between subjective ratings and performance is highlighted.

---

* Now at; NASA - Ames Res Ctr, MS 239-3, Moffet Field, CA 94035

PRECEDING PAGE BLANK NOT FILMED

INTRODUCTION

Possibly the most enduring issue in experimental psychology concerns the question of whether people are really aware of what happens in their heads. For many of the earliest American psychologists such as James, Titchner, or Angell it seemed obvious that introspection would provide much of the science's data. But, during the heyday of behaviorism it was generally accepted that only observable behavior was worth studying. While few researchers subscribe to, or espouse such strong anti-mentalistic claims today there is still considerable evidence of a debate between those who believe subjects can provide useful information about what is going on in their heads and those who do not. For example, the debate between Nisbett and Wilson (1977) and Ericsson and Simon (1980) centers on just this issue.

This question is not only of philosophical interest there are some serious applied consequences for practicing human factors personnel as well. Most obvious, is the question of whether subjective assessments of workload posses any value in system evaluation? If, as Ericsson and Simon's work would suggest, subjects are capable of accurately reporting whether parts of their internal cognitive processing equipment are being over-worked, then the subjects' workload assessments would deserve a valued place in the evaluation of man-machine interfaces. On the other hand, if Nisbett and Wilson are correct is asserting that subjective assessments are actually post hoc logical analyses, then human factors practitioners might be better served by creating their own logical assessment algorithms for generating workload estimates.

The first step in resolving this debate is to see if subjective assessments of workload can accurately predict performance. This would at least establish the possibility that they are based on legitimate introspection of the workings of the mind and may be of use to human factors personnel. If the workload assessments have absolutely no relationship to performance, then the question of the value of introspection might remain a topic of debate for those of a purely theoretical bent, but it would be a sterile issue for applied personnel.

The present study further investigates these issues. In this study a standard compensatory tracking task was utilized. The difficulty of the tracking was manipulated in two different fashions. One manipulation involved increasing the control order of the system, while the other increased the bandwidth of the disturbance forcing function. These two different methods were selected to emphasize different aspects of increasing difficulty. The order manipulation is expected to cause a relatively substantial increase in the early perceptual/central processing stages of information processing, because dealing with the increased control order involves use of a more sophisticated strategy including more anticipation. In contrast, increasing bandwidth speeds up everything, but does not require a qualitatively different strategy. Therefore, relative to the order manipulation, the bandwidth manipulation's increasing difficulty might be considered mostly due to increased response processing demands (Wickens, Gill, Kramer, Ross &

Donchin, 1981).

Since, Ericsson and Simon have suggested that verbal reports are most sensitive to information heeded in primary memory, it would follow that subjective workload assessments would be more accurate in the case of increasing difficulty by increasing control order.

To investigate this hypothesis an analysis of covariance (ANACOVA) procedure was employed. The procedure assumes that some portion of the error component of the dependent measure is predictable if the subject's score on some related measure, say "X", is known. If so, then the ANACOVA procedure allows a researcher to partial out subject and/or task differences in the "X" data from the dependent measure scores. Therefore, to the degree that the workload ratings are based on information concerning cognitive processes effective in influencing performance quality the use of the ratings covariate in the ANACOVA technique should attenuate or eliminate the effect of task difficulty in the present study's analysis. Consequently, based on the experimental hypotheses, it would be predicted that the ANACOVA procedure would be more effective in the analysis of the increasing control order data than in the analysis of the increasing bandwidth data.

(For a general discussion of the ANACOVA technique refer to Meyers (1979); for an example of its application to engineering psychology research see Ackerman and Wickens (1982).)

The same basic methodology can be applied to another somewhat related topic; i.e., the question of a subject's ability to voluntarily control the amount of effort applied in performing a task. In this case, using the subject's ratings of effort as a predictor of performance should wipe out any effects of lowered effort on the performance scores.

Also, comparison of the results of single-task effort manipulation to dual-task biasing could have important implications as well.

In summary, the intent of this study was to test the following hyupotheses:

(1) Subjects have access to subjectively available data concerning the processing of task related information. This data can be collected via verbal reports.

(2) The closeness of the relationship between verbal reports and performance is influenced by the type of processing employed by a task and can be tested by use of an ANACOVA procedure.

(3) The same general procedure can be used to explore subjects' abilities to voluntarily control effort.

METHOD

Subjects

Nine right-handed students of the University of Illinois (7 male, 2 female) participated in the study. All subjects had normal or corrected vision and were paid $3.50 per hour for their participation.

Apparatus

Subjects were seated in a light and sound attenuated booth. Stimuli were displayed on a 10 cm x 8 cm Hewlet-Packard Model 1330a cathode ray tube (CRT) positioned approximately 90 cm in front of the subject and slightly below eye level. Responses were collected via a Measurement Systems Incorporated Model 435 spring-centered, joystick affixed to the left armrest of the subject's chair. A Raytheon 704 sixteen bit digital computer with 24K memory was used to generate the experimental displays and record subject performance.

Tracking Task

A single-axis compensatory tracking task required subjects to null an error indicated by a lateral displacement of a circular cursor from a target line located in the center of the CRT display. The cursor was controlled by right-left movements of the joystick. The task was driven by a band-limited Gaussian disturbance input with an upper cutoff frequency which could be set at 0.3, 0.4, 0.5, or 0.6 Hz. The system control dynamics could be set as a pure first-order velocity control (order = 1.0), a pure second-order acceleration control (order = 2.0) or a linear combination of parts of first and second order control (order = 1.5).

Six different combinations of disturbance input frequency (bandwidth) and system control dynamics (order) were utilized in the experiment. The easiest condition, which will be referred to as the "standard", combined the lowest bandwidth (i.e., 0.3 Hz) with the easiest control order (i.e., first order). Higher levels of difficulty were then achieved by increasing either bandwidth or order, but never both. Three higher difficulty conditions were formed by combining first order control with 0.4, 0.5, or 0.6 Hz. Additionally, two more conditions were generated by combining 0.3 Hz bandwidth with either 1.5 or 2.0 order system dynamics.

Experimental Design

Two experiments were performed with the subjects in this study: (1) a subjective assessment of workload experiment, and (2) a voluntary control of effort experiment.

In the subjective workload assessment experiment a relative rating procedure was used. The standard condition was arbitrarily given a workload rating of 10. All other test conditions were compared to the standard condition by first giving the subject a 15 second exposure to the standard follwed by a full two minute trial of the test condition. After completing the trial the subject would rate the test condition relative to the standard condition; twice as much workload would be a rating of 20, half as much workload would be a rating of 5, and so on. Within each block of the workload assessment experiment one full two-minute trial of the standard conditon would be run. The subject would be warned prior to this trial that this "standard-alone" condition was being run merely to obtain a performance score and that there was no reason to rate it. However, within each block there was also a "trick" test which consisted of the 15 second exposure to the standard being followed by a full two-minute test trial of the standard condition. Although this experiment is designed primarily as an investigation into the subjects' ability to identify differences in workload, this trick-standard was included as a test of the subjects' ability to identify a "same" condition. Also, the trick-standard was included as a hedge against possible motivational differences contaminating the comparison of performance in the standard alone condition to performance in the other test conditions. Each block, therefore, consisted of 6 test-condition trials which were rated by the subject and 1 standard-alone trial. Each subject performed two complete blocks of the subjective assessment experiment.

In the voluntary control of effort experiment each trial consisted of a set of three connected two-minute trials of one of the six tracking conditions. Prior to the first trial of the set the subjects were instructed to perform as well as they could (i.e., 100% effort). Just before starting the second trial of the set subjects were told to hold back a little and to try to work at approximately a 70% effort level. Between the second and third trial of the set subjects rated how much effort they expended on the previous trial and were then instructed to hold back even more and attempt to perform at approximately a 30% level of effort. Following the third trial of the set, subjects once again rated how much effort they expended on the previous trial. Each block of the voluntary control of effort consisted of six sets of three connected two-minute trials; one for each tracking condition. Each subject performed two blocks.

In both experiments each subject received a unique random order for each block of trials.

Procedure

The experiment consisted of five 1-hour sessions for each subject. The first session consisted of instructions followed by 18 practice trials. Following each trial performance feedback was delivered and, if appropriate, suggestions for improving performance were given. The second session was also a practice session which opened with four trials using the subjective workload assessment procedure described above. This was followed by four sets of three trials using the voluntary control of effort procedure. The actual experiment started in the third session which consisted of one block of the subjective assessment of workload experiment, followed by three sets of three trials of the voluntary control experiment. The fourth session consisted of six sets of three trials of the voluntary control experiment. (The first three sets completed the first block and the second three sets started the second block). The final session started with the second block of the subjective workload assessment experiment and finished with the remaining three sets of trials for the voluntary control experiment.

RESULTS

A single analysis strategy will well be employed in analyzing the results of both experiments. First, the data for the two different methods of increasing difficulty (i.e., order and bandwidth manipulations) will be separated for individual analysis. In each case the analysis will start with an investigation of the effects of tracking performance, as reflected in Root Mean Square Error (RMSE). Then an analysis of the effects of the task conditions on subjects' ratings will be performed. (In Experiment 1 these will be ratings of workload, while in Experiment 2 they will be effort ratings.) Following this, the ratings data will be utilized as a covariant in a ANACOVA with the performance data as the dependent measure. This should serve as a test of the validity of using the ratings in explaining variability in performance. If ratings tap a legitimate source of information it would be expected that the covariance analysis technique would eliminate, or at least substantially reduce, performance effects due to increasing difficulty in the first experiment or due to decreasing effort in the second experiment.

Subjective Ratings Experiment

Before moving into the basic analysis procedure a quick look at the data for the two "standard" conditions are in order. First, there is the question of whether the uniqueness of the standard-alone condition caused any performance differences relative to the trick-standard. Although there was a small standard-alone advantage indicated in the RMSE means (0.158 vs. 0.166), a t-test comparing these two conditions failed to reach significance ($t(8) = 1.078$, $p > 0.05$). Another t-test was performed to see whether the mean workload ratings for the trick-standard were different from 10. Subjects exhibited a slight tendency to underestimate the difficulty of the trick-standard relative to the arbitrary level of 10 (mean rating = 9.4). However, this difference also failed to reach significance ($t(8) = -0.662$, $p > 0.05$). Since only two of the nine subjects would admit during debriefing to having been suspicious of the condition, this finding supports the conclusion that subjects were reasonably adept at identifying equivalent levels of workload. In the following analyses the trick-standard condition will provide the data used as the standard condition baseline.

Moving on to the question of the effects of the difficulty manipulations; Figure 1 displays the data for both the order manipulation (in the left column) and the bandwidth manipulation (in the right column). The two graphs in the top row display the RMSE

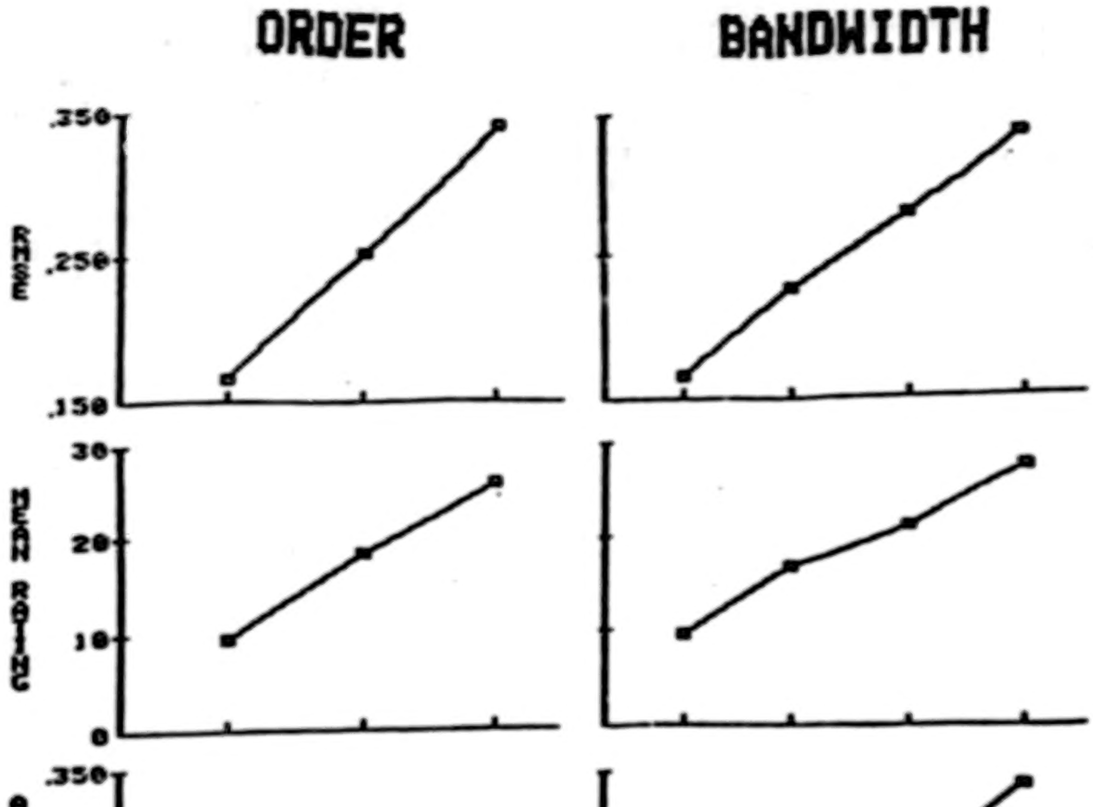


Figure 1. Illustration of the data from the subjective workload assessment experiment.

performance for the different difficulty conditions collapsed over block. (Note that the leftmost point in all of the graphs represents the data for the same standard condition.) These data for each difficulty manipulation were subjected to a Difficulty x Block ANOVA. difficulty had a significant effect in both the order (F (2,16) = 38.16, $p < 0.0001$) and the bandwidth manipulations (F (3,24) = 102.32, $p < 0.0001$). As expected, in both cases performance worsened in the conditions designed to be more difficult. Block had a significant effect only in the bandwidth manipulation condition (F (1,8) = 7.87, $p < 0.0230$). Subjects showed a slight improvement from the first to second block.

In the second row of the figure the ratings data are displayed. These data were also analyzed by a Difficulty x Block ANOVA. In both these analysis difficulty was again found to have a significant effect (F (2,16) = 47.26, $p < 0.0001$ in the order analysis; F (3,24) = 32.89, $p < 0.0001$ in bandwidth analysis). Not surprisingly subjects rated the more difficult conditions as incurring greater workload. Block was not significant in either analysis.

The bottom row displays the adjusted means for the ANACOVA analyses of the RMSE data with the ratings used as the covariate. These adjusted means are part of the output of the ANACOVA analysis and represent the resulting dependent measure scores when differences in the covariate are partialled out. On the left side it can be seen that the covariance technique is very effective in eliminating the difficulty effect of the order manipulation. This is indicated by the pronounced flattening of the curve relative to the corresponding graph in the top row. In other words, the performance differences are explained by the differences in the subjective workload ratings. In contrast, the adjusted means associated with the bandwidth manipulation show virtually no change relative to the raw performance scores. In the bandwidth ANACOVA both difficulty and block remained significant effects (F (3,23) = 18.36, $p < 0.0001$ and F (1,7) = 8.03, $p < 0.0253$, respectively).

To summarize, the overall pattern of results in the first study is completely in keeping with the predictions. The subjective workload assessments are closely related to performance in the case of increasing control order, but not in the case of increasing bandwidth.

Voluntary Control of Effort Experiment

Figure 2 displays the data from the analyses of the voluntary control experiment. The data from the 100% effort level were not included in the analyses because there were no corresponding ratings.

The top two graphs display the raw performance effects. These data were subjected to a set of Difficulty x Instructed Effort x Block ANOVAs. Increasing difficulty increased RMSE in both the order (F (2,16) = 33.98, $p < 0.0001$) and the bandwidth (F (3,24) = 22.3, $p < 0.0002$) analyses. Also, in both the order and bandwidth analyses lowering the instructed effort level from 70% to 30% significantly

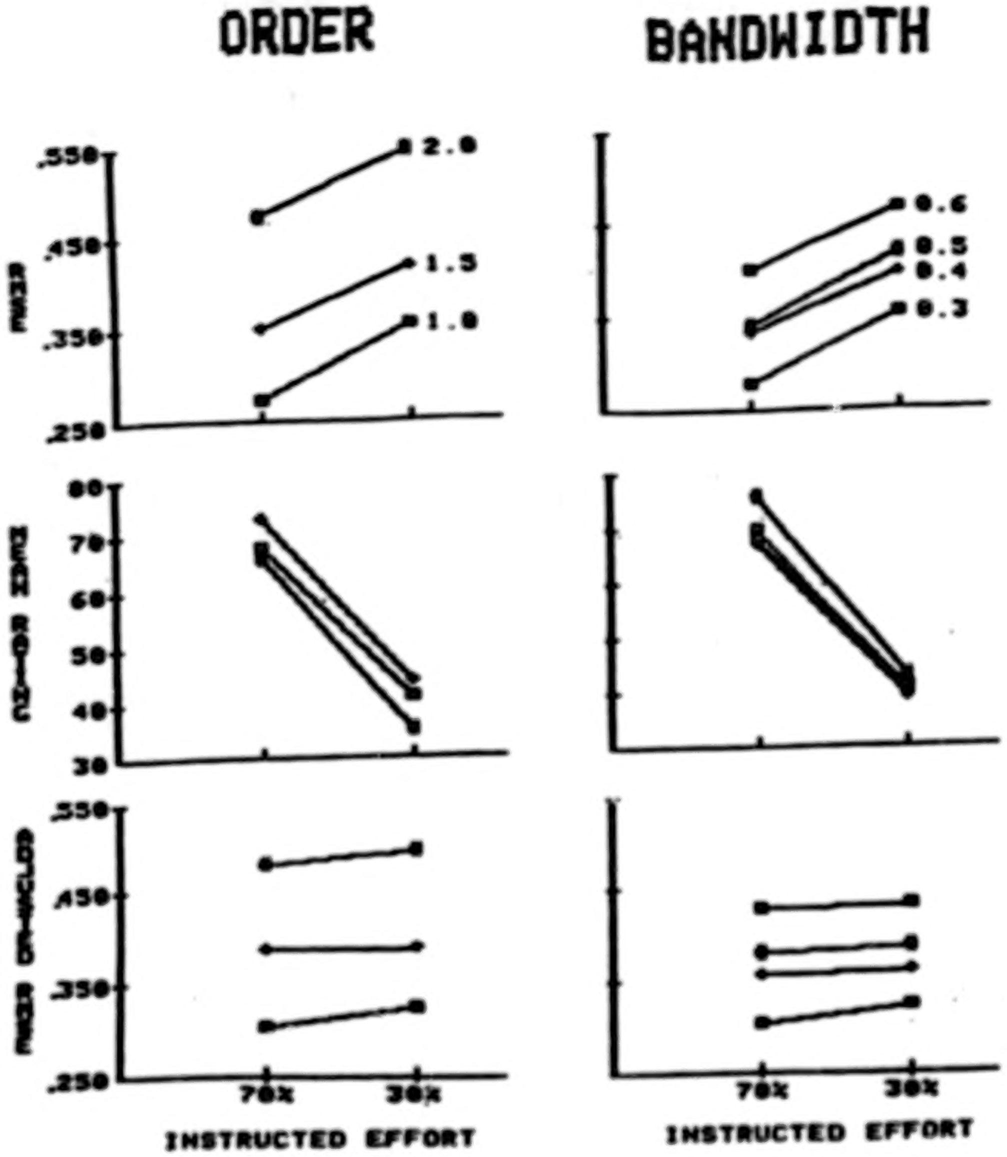


Figure 2. Illustration of data from the voluntary control of effort experiment.

increased RMSE ($\underline{F}$ (1,8) = 34.30, $\underline{p} < 0.0004$ and $\underline{F}$ (1,8) = 35.50, $\underline{p} < 0.0003$, respectively). There were no significant effects due to block, nor was there any significant interaction.

The middle row of graphs display the data from the corresponding set of analyses performed on the subjects' ratings of effort. The ANOVAs performed on this data detected a significant effect of instructed effort in both the order ($\underline{F}$ (1,8) = 87.26, $\underline{p} < 0.0001$) and bandwidth ($\underline{F}$ (1,8) = 62.63, $\underline{p} < 0.0001$) analyses. In both cases as subjects were instructed to expend less effort they responded with lower estimates of effort expended.

The bottom two graphs display the adjusted means from the output of the ANACOVA analyses. In both analyses there is a pronounced reduction in the slopes of the lines, resulting from the elimination of the previous significant effect of the instructed effort variable. In both analyses the difficulty variable remained a significant effect ($\underline{F}$ (2,15) = 34.54, $\underline{p} < 0.0001$ in the order analysis; $\underline{F}$ (3,23) = 30.25, $\underline{p} < 0.0001$ in the bandwidth analysis).

In general then, the results of the analyses in this experiment are even stronger than those of the last experiment in demonstrating that the use of subjective ratings as a covariate can eliminate previously significant effects in the analysis of performance data. Further, it demonstrates that this can be accomplished while leaving the effect of other variables in the analysis untouched (in this case the difficulty variable).

Single-Task Effort vs. Dual-Task Bias

The effect of effort on performance has most often been investigated in a dual-task environment. This is done by asking subjects to favor one task over another to some degree. Subjects seem to be quite capable of biasing their performance in such a manner; although it has been suggested that subjects can only achieve three to five gross levels, at best (Navon, 1984). Given the central role the concept of effort plays in most capacity notions of attention close scrutiny of the empirical effects of laboratory manipulations of effort seems essential. The present study manipulated effort in a single-task environment with a tracking task that had previously been used in a dual-task biasing experiment (i.e., Vidulich & Wickens, 1981). The logic of comparing the results of the two studies was inescapably compelling.

To review, briefly, Vidulich and Wickens (1981) manipulated bias in a dual-task environment in which a first or second order tracking task was performed concurrently with a Sternberg memory search task. The bandwidth of the tracking task was always 0.3 Hz. The Sternberg task was tested with four different input/output (i/o) configurations; auditory/speech (A/S), auditory/manual (A/M) visual/speech (V/S), visual/manual (V/M). Subjects were asked on different trials to favor one or the other task by a 70% to 30% priority split. Results indicated that the bias variable significantly influenced performance in both tasks. Also, as predicted by multiple resource theory, biasing was most potent when competition for resources was highest; that is, when the tracking (which is a V/M task) was concurrently performed with the V/M Sternberg task.

Consequently, the data from the V/M Sternberg dual-task trials were selected for comparison to the corresponding first and second order tracking conditions of the present study. Since the data from the Vidulich and Wickens (1981) study were in the form of decrement scores (i.e., the difference between single-task and dual-task performance) it was necessary to convert the present data to a comparable format. Decrement scores for the present study were generated by subtracting the 100% effort trials RMSE from the RMSE on the corresponding 70% effort and 30% effort trials.

The data from both experiments were then subjected to an Experiment x Control Order x Effort ANOVA. The mean RMSE decrements for both experiments is displayed in Figure 3. There was a significant main effect of effort ($\underline{F}$ (1,16) = 37.42, $\underline{p} < 0.0001$). Reducing effort (or lowering the bias) increased decrements in both experiments. There was a significant effect of experimental procedure on decrement size ($\underline{F}$ (1,16) = 11.89, $\underline{p} < 0.0033$). Overall, the size of decrements were larger in the single-task experiment than in the dual-task experiment. But, most important of all, there was a significant interaction between the experiment and the effect of manipulating effort ($\underline{F}$ (1,16) = 11.01, $\underline{p} < 0.0043$). In the single-task case changes in effort were much more potent than dual-task changes in bias.

The ramifications of these findings on hypotheses of effort control mechanisms will be reviewed later.

DISCUSSION

In the most general sense the present study can be considered an investigation into the value of phenomenal events in engineering psychology research. This general issue was approached in two distinct manners: Firstly, the strengths and weaknesses of individuals' assessments of their internal states as workload measures were studied. Secondly, the subjects' abilities to voluntarily control the effort expended in performing a single task was investigated. In general, the results of both studies indicate an encouraging facility on the part of the subjects in both assessing and controlling the internal processing related to task performance. Yet, the two studies also offer some cautions regarding the limitations of these abilities.

In the workload assessment part of the study a preliminary viewing of the data might be more encouraging than is warranted. Both difficulty manipulations successfully influenced performance in the expected directions and analysis of the workload ratings seems at first glance to closely reflect the same trends. However, when the ANACOVA procedure is used the ratings can only successfully explain the

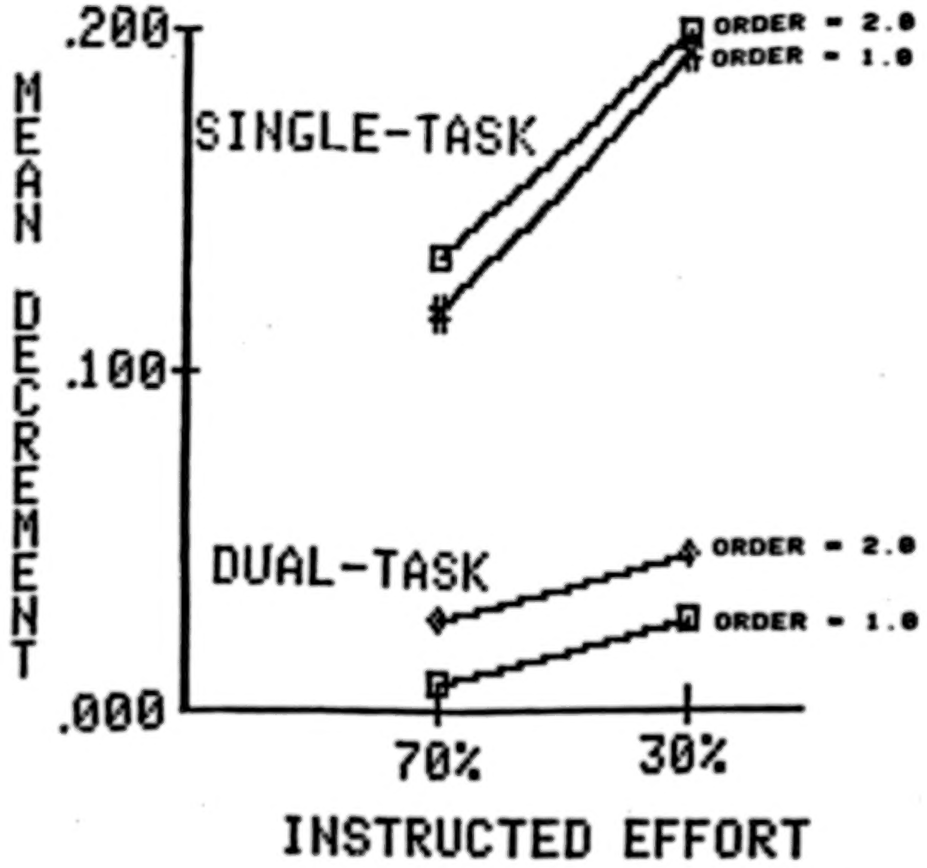


Figure 3. Comparison of single-task voluntary control of effort to dual-task biasing (from, Vidulich & Wickens, 1981).

performance in the case of the control order manipulation. In the bandwidth manipulation ANACOVA, the effect remains significant and the adjusted means show virtually no change relative to the raw performance means. This result, though damaging to the casual use of subjective assessments of workload, is entirely in keeping with both Ericsson and Simon's (1980) theoretical interpretation of verbal report validity and with empirical demonstrations of the locus of effects of the two difficult manipulations.

Ericsson and Simon propose that verbal reports are most sensitive to events in short-term memory. Within Wickens (1980) multiple resource model this would imply that verbal reports are most effective in assessing the state of perceptual/cognitive processing as opposed to response execution processing. Previous research (i.e., Wickens, Gill, Kramer, Ross & Dochin, 1981; Wickens & Derrick, 1981) has indicated that the effect of increasing tracking control order is heavily biased towards increasing perceptual/cognitive processing demands relative to response execution demand. In contrast, the same research indicated that increasing difficulty by increasing the forcing function bandwidth of the tracking is somewhat biased towards increasing the response execution processing demands more than the perceptual/cognitive processing demands.

Therefore, in the case of the order manipulation the subjects' subjective assessments are based on the same locus of processing as is most heavily influenced by the difficulty manipulation. This results in a close relationship between the ratings and the performance in the control order manipulation data. But, in the bandwidth manipulation the increasing difficulty has a major portion of its effect isolated in the response execution processing which is relatively inaccessible by subjective assessments. In this case, since the subjective assessments are not based on the same processing that is responsible for the difficulty effect there is not such a intimate relationship between ratings and performance and the ANACOVA analysis cannot eliminate the significant difficulty effect.

In summary, the results of the first study favor an extension of Ericsson and Simon viewpoint to the area of subjective workload assessment. Such assessments will tend to be most accurate when the phenomena of interest is isolated in the perceptual/cognitive processing as opposed to the response execution processing. This identifies a limit of the applicability of subjective workload assessments, but it is a limitation which can be lived with. Fortunately, most contemporary applications of workload assessment are concerned with the operator's decision making processes and these processes should be accessible by verbal reports.

The second part of the study investigated the voluntary control of effort in single-task tracking at different levels of difficulty. In this study the effects were virtually identical for the two difficulty manipulations. In both cases, the tracking error increased as the instructed level of effort was reduced, the ratings of effort were consistent with the instructed effort level, and in the ANACOVA analysis the ratings covariate eliminated the effect of the instructed level of effort from the adjusted performance means.

Generally, these results can be interpreted as implying that subjects are quite capable of adjusting their effort in a reasonable and consistent manner. However, the results encourage speculation on the underlying mechanism of effort control and imply that this mechanism might be fairly complex.

First, the fact that there is no interaction in the performance data between effort and level of difficulty is intriguing, especially in the order manipulation data. (Note that this lack of a Effort x Difficulty Level interaction persists even if the performance data from the 100% effort trials are included.) If the increasing control order requires the use of a more complex mental model, as suggested by Wickens et al. (1981), then we would expect that an initial reduction in effort might have a somewhat greater effect in the higher control order data. In other words, the figure of the performance data for the different levels of control order plotted accross effort level might be expected to give a "fan-like" appearance as opposed to the set of parallel lines which actually appears in the top portion of Figure 2. Since this interaction does not appear, it calls into question the nature of the effort control mechanism. Possibly, rather than controlling level of effort directly subjects employed an indirect strategy in which effort is "controlled" by loosening the performance criterion they attempt to achieve. Such a mechanism would not be expected to produce a Effort x Difficulty Level interaction.

Another question is raised by the comparison of the present study's results to the results of biasing in a dual-task environment. The finding that the effect of controlling effort in a single-task condition is more potent than controlling dual-task biasing initially appears to be explicable within a multiple resource framework. Previous research (Vidulich & Wickens, 1981) suggested that the size of the bias effect in a dual-task experiment increases as the amount of resources competed for by the two tasks increases. This implied that in the dual-task case differential effort only occured in those resource pools from which both tasks demanded resources. In the single-task case subjects might adopt an across-the-board effort reduction affecting all resource pools. This mechanism appears to explain the comparison results, but does not seem to be easily reconciliable with the possibility of a "criterion relaxation" mechanism of effort control as discussed previously.

Taken as whole, the results of the second study indicate that subjects can voluntarily control their effort in a single-task environment. But, the mechanism by which they do this remains an open question and generalization from single-task effects to dual-task environments appears untenable, at this time.

Both studies in this investigation agree on at least one point; namely that the ANACOVA procedure appears to be a useful statistical technique for investigating the relationship between ratings and performance. This is especially true in the first experiment where the raw performance and ratings data showed virtually the same trends for both difficulty manipulations, but the ANACOVA procedure showed that in the bandwidth data the ratings were not closely enough associated with the performance to expain it. The ANACOVA procedure appears to have great potential as a tool in engineering psychology research.

REFERENCES

Ackerman, P. L. & Wickens, C. D. (1982). Methodology and the use of dual and complex-task paradigms in human factors research. In R. E. Edwards (Ed.), Proceedings of the Human Factors Society 26th Annual Meeting (pp. 354-358). Santa Monica: Human Factors Society.

Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. Psychological Review, 87, 215-251.

Meyers, J. L. (1979). Fundamentals of experimental design. (3rd ed.) Boston: Allyn & Bacon.

Navon, D. (1984). Resourses - A theoretical soup-stone? Psychological Review, 91, 216-234.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

Vidulich, M. & Wickens, C. D. (1981). Time-sharing manual control and memory search: The joint effects of input and output modality competition, priorities, and control order (Tech. Rep. EPL-81-4/ONR-81-4). Urbana: University of Illinois, Engineering-Psychology Laboratory.

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), Attention and Performance VIII, Hillside, NJ: Erlbaum.

Wickens, C. D. & Derrick, W. (1981). The processing demands of higher order manual control: Application of additive factors methodlogy (Tech. Rep. EPL-81-1/ONR-81-1). Urbana: University of Illinois, Engineering-Psychology Laboratory.

Wickens, C., Gill, R., Kramer, A., Ross, W., & Donchin, E. (1981) The cognitive demands of second order manual control: Applications of the event related potential. Proceedings of the Seventeenth Annual Conference on Manual Control, JPL Publication # 81-95, 7-17.

N85
14541

UNCLAS

N85 14541 D6

DECISION TREE RATING SCALES FOR WORKLOAD ESTIMATION: THEME AND VARIATIONS

Walter W. Wierwille
Julie H. Skipper
Christine A. Rieger*

Vehicle Simulation Laboratory
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

SUMMARY

The Modified Cooper-Harper (MCH) scale has been shown to be a sensitive indicator of workload in several different types of aircrew tasks (Wierwille and Casali, 1983). The study to be described in this paper was undertaken to determine if certain variations of the scale might provide even greater sensitivity and to determine the reasons for the sensitivity of the scale. The MCH scale, which is a 10 point scale, and five newly devised scales were examined in two different aircraft simulator experiments in which pilot loading was treated as an independent variable. The five scales included a 15 point scale, computerized versions of the MCH and 15 point scales, a scale in which the decision tree was removed, and one in which a 15 point left-to-right format was used.

The results of the study indicate that while one of the new scales may be more sensitive in a given experiment, task dependency is a problem. The MCH scale on the other hand exhibits consistent sensitivity and remains the scale recommended for general use. The MCH scale results are consistent with earlier experiments also. This paper presents the results of the rating scale experiments and also describes the questionnaire results which were directed at obtaining a better understanding of the reasons for the relative sensitivity of the MCH scale and its variations.

INTRODUCTION

It has gradually become recognized that rating scales, properly designed and tested, represent a sensitive and economical means for estimating mental workload. They can be used in a systematic manner to obtain a single numerical response, which estimates the magnitude of the multidimensional construct of mental workload.

One of the most popular and widely accepted scales is the so-called Cooper-Harper scale (Cooper and Harper, 1969). This scale incorporates an unusual decision tree and descriptors directed at handling qualities, stability, and workload. The scale is well suited for estimation of workload in manual control systems. For example, Wierwille and Connor (1983) showed that the scale was quite sensitive to changes in turbulence level and longitudinal stability in an instrument landing task. Variations of the original scale have also appeared, but they too have been directed primarily

---

* Now with Hughes Aircraft Co., Ground Systems Group, Fullerton, CA

PRECEDING PAGE BLANK NOT FILMED

toward manual control applications (North and Graffunder, 1979; O'Connor and Buede, 1977; Siefert, Daniels, and Schmidt, 1972; and Wolfe, 1982). More recently, Wierwille developed a modication of the scale, called the Modified Cooper-Harper (MCH), which could be universally applied in mental workload estimation, regardless of the type of loading imposed by the task (Wierwille and Casali, 1983)*. In particular, the scale was designed to provide a global measure of mental workload in tasks having loading along communications, mediational, and perceptual dimensions. The scale was subsequently tested and found to be experimentally sensitive and valid in three independent simulator experiments.

Because the MCH scale had already been tested and found adequate, questions could be asked regarding the reasons for its sensitivity and regarding improvements that might be made. Thus, another study was undertaken in which the MCH scale was systematically varied in an effort to gain greater insight. Specifically, the MCH scale and five variations emphasizing major design aspects were used in this study. The six rating scales were then used in two different experiments, one involving mediational (cognitive) loading and one involving communications loading. The results are reported in this paper.

METHOD

Thirty six pilots (30 private and 6 student) participated, each participating in both experiments. Four pilots were females, and 32 were males. The pilots were tested for hearing and vision using standard tests. They were paid for their participation.

The aircraft simulator used for the two flight task experiments was a modified Singer-Link GAT-1B moving base, simulator. The simulator had three degrees of physical motion--yaw, pitch, and roll. For both experiments, the simulator was equipped with translucent blinders to eliminate outside distractions. The ambient illumination was held constant. A lapel microphone and speaker system were installed in the simulator cockpit so that the subjects could communicate with the "tower" (experimenter). To assure that the subjects were continually providing input control to the simulator, mild, random wind gusts were introduced into the simulator flight dynamics. For the mediational experiment the simulator was additionally equipped with a Kodak Ektagraphic slide projector (Model 260) mounted in front of the simulator windscreen. To computerize two of the six rating scales, a TRS-80 Model III micro-computer was used. The rating scales were programmed in BASIC, and the subject ratings were performed on the TRS-80 computer in a reduced glare setting.

Six rating scale designs were used in both the communications and the mediational experiments. The first rating scale was the Modified Cooper-Harper (MCH) rating scale described earlier. The MCH scale has a 3-3-3-1, decision tree scale structure. The second rating scale, COMPMCH, was a computerized version of the MCH scale. The TRS-80 was used to administer the MCH scale to the subjects on a decision-by-decision basis. The subjects were only permitted to deal with one primary decision at a time. Thus, the subjects did not know where each primary decision would lead on the rating scale. (A typical computer frame of the COMPMCH scale is illustrated in Figure 1). The computer implemented scale was used to discover whether or not the decision tree logic of the MCH scale was being utilized or if the subjects were merely rating on the basis of the category descriptors and numerical values. After each computer rating, the subjects were asked by the computer if they were satisfied with their rating. If they were not satisfied, the program repeated the procedure for rating. When the subjects were satisfied with their rating, the rating value was recorded. To investigate the possibility of additional rating scale categories increasing the sensitivity of the MCH scale, the third rating scale, MCH+ (Figure 2), expanded the MCH scale to a 15 point decision-tree rating scale. One additional category was added to the first three rating groups and two additional categories were added to the last rating group, giving a 4-4-4-3 scale structure. The COMPMCH+ scale, the fourth rating scale, was a computerized version of the MCH+ scale and was implemented in the same manner as the COMPMCH scale.

---

* Figure 1 of Wierwille and Casali (1983) shows the MCH scale.

In the fifth rating scale, the PBMCH (performance-based MCH) scale (Figure 3), the primary decision hierarchy was changed by manipulating the tree structure. The PBMCH decision tree flow was from left to right and the first decision was concerned with the errors of the subjects in performing the instructed task. This scale was used in an attempt to improve the sensitivity of the MCH scale by modifying the decision tree logic of the scale requiring an assessment of the subjects' errors first in the rating process. Finally, the sixth rating scale, the NDT (no decision tree) scale (Figure 4), removed the visual decision tree structure from the MCH scale to find out how the visual tree affected the sensitivity of the MCH scale. The NDT scale presents the MCH rating information in a tabular format.

Identical experimental designs were used in both the communications and mediational experiments. Data were analyzed as a rating scale by load (6x3) design. Load presentation order was completely counterbalanced. Each subject used only one rating scale, which was the same scale for both experiments. Six subjects used each scale, resulting in a total of 36 subjects. Thus, rating scale was a fixed-effects between-subjects variable and load level was a fixed-effects within-subject variable. Experience level was controlled by dividing the 36 subjects into sextiles according to flight hours and then selecting one subject from each sextile for each rating scale.

COMMUNICATIONS EXPERIMENT

The communications experiment task and protocol were identical to those used by Casali and Wierwille (1983) in an experiment comparing many different kinds of workload estimation techniques. The reader is referred to this earlier experiment for a detailed description of the task. Briefly, the aircraft control and communications requirements were performed simultaneously in the task. After reaching altitude, subjects maintained straight and level flight in mild turbulence until instructed to make changes.

For the communications aspect, the subjects listened to an 8-minute tape recorded message that was played over the cockpit speaker system. The taped communications scenario was a "tower" controller with a male voice. The subjects were required to attend to two components of the taped scenario. The first component consisted of pilot commands. In the commands, the subjects were asked to change and report aircraft parameters (e.g. change altitude, heading, and radio frequency, and report airspeed, aircraft model, altitude, and heading). In the second component of the taped scenario, the subjects were presented with strings of randomly constructed aircraft call signs. Each call sign consisted of two international phonetic letters and two single digits (e.g. Alpha-Four-Bravo-One). Out of the randomly presented call signs the subjects were instructed to respond "now" to their specific call sign "One-Four-India-Echo" and to any of 5 permutations of the call sign which always featured "one" in the first position of the call sign. Thus, the subjects had six target call signs to listen for, each beginning with "one", as a cue to listen to what followed.

The communications load was varied in this experiment by manipulating the presentation rate of the target call signs and the non-target permutations of "One-Four-India-Echo." The three load levels were: low, 1 target every 12 seconds with 0 non-target permutations; medium, 1 target every 5 seconds with 30% permutations; and high, 1 target every 2 seconds with 40% permutations.

The experiment began with a practice flight which contained equal portions of all three communications load levels. The data run flights then followed —one at each load level. After each of the experimental flights, the simulator was placed in autopilot control and the subjects left the simulator to make a rating on their respective rating scale. They then completed a questionnaire. The questionnaire was administered to allow the subjects to describe the factors on which their ratings were based. After the final experimental flight the subjects landed the simulator and were dismissed. (They returned later the same day to participate in the mediational experiment. After completion of both experiments, they were debriefed, paid, and dismissed.)

In addition to the ratings, all verbal responses of the subjects were recorded and later scored for errors of omission, errors of commission, and reaction times.

COMMUNICATIONS EXPERIMENT RESULTS

The main statistical analysis results for the communication experiment are presented in Table 1. The rating scale scores for each rating scale were first subjected to a one-way analysis of variance. An α-level of 0.01 was specified to account for the fact that six different rating scale ANOVA's were performed. Mean values, in terms of Z-scores for each rating scale, were also computed and appear in the table. For those ANOVAs resulting in significance at $p < 0.01$, Duncan's multiple comparisons were carried out.

The results of the tests indicate that the MCH, COMPMCH, and PBMCH scales resulted in significant ANOVA's. All three scales increased monotonically with load. Furthermore, the three scales exhibited similar sensitivity, with the MCH showing slightly greater sensitivity than the other two.

Two multivariate analyses were performed on the voice response measures, at $\alpha = 0.05$, to test the data for performance variations due to either the rating scale groups or the pilot experience level groups. The three measures used were errors of omission, errors of commission, and response times. The Wilk's $\underline{U}$-likelihood ratio statistic $\underline{F}$-approximation is reported. The results showed that there were no statistically significant performance variations among the rating scale groups ($\underline{F}$ (15, 77) = 1.40, $\underline{p} = 0.1663$) nor the pilot experience level groups ($\underline{F}$ (15, 77) = 1.61, $\underline{p} = 0.0895$).

To obtain general information regarding the effects of pilot experience level and load presentation order on the ratings of the subjects, converted and collapsed raw score data were analyzed in two separate ANOVAs. The results indicated that neither the pilots' experience levels ($\underline{F}$ (5, 30) = 2.43, $\underline{p} = 0.0579$) nor the load presentation orders ($\underline{F}$ (1, 35) = 0.43, $\underline{p} = 0.5173$) affected the ratings of the pilots. The experience level results were analyzed further using regression, but the additional analyses did not provide significant findings.

The responses to the questionnaire presented to the subjects indicated a shift in tone from positive to negative as the load levels progressed from low to medium to high. A Chi-square analysis on a 2 x 3 contingency table, response type by load levels, resulted in $\chi^2_2 = 68.326$, $\underline{p} < 0.0001$, confirming the change of tone due to load in the responses of the subjects. Typical response classifications were "time-sharing", "aircraft control", and "recognition of target call signs".

Finally, it is worth mentioning that the MCH scale results in this experiment were virtually identical to the MCH scale results obtained in the earlier (Casali and Wierwille, 1983) study. This indicates a high degree of repeatability for the MCH scale.

MEDIATIONAL EXPERIMENT

This experimental task and protocol were also identical to an earlier experiment in which mediational activity was emphasized (Wierwille, Rahimi, and Casali, 1984) and in which many different workload techniques were evaluated. The reader is referred to this earlier experiment for a detailed description of the task. Briefly, the overall task consisted of two components: straight and level flight in mild turbulence (within specified tolerances), and solution of navigation problems. Subjects performed the tasks simultaneously with instructions indicating equal priority.

The navigation task of solving wind triangle problems was used to interject mediational loading into the basic flight task. Wind vector triangles depicted on slides involved solving for the effects of wind direction and velocity on the path and speed of an aircraft. The slides contained both a problem triangle and a reference triangle. The reference triangle provided numerical values associated with the triangle legs and the angles corresponding to the problem triangle.

The difficulty of the navigation problems was manipulated by varying the question type, the numbers used in the mental calculation of the problems, and the orientation of the reference triangles. Depending upon the question type, the problems required triangle comparison, triangle comparison followed by an addition or subtraction, or triangle comparison followed by an addition or subtraction and a subsequent division. For all load levels, the slide presentation rate was held constant at a rate of one slide per 25 seconds. Subjects expressed their answers verbally. These responses were recorded for later use in computing response time and number of correct responses. It is important to note that the subjects did not implement the solutions to the navigation problems. They maintained constant altitude, heading, and airspeed throughout each flight.

The general flight procedures for the mediational experiment were the same as for the communications experiment. In particular, one practice and three data flights were performed, and subjects left the simulator while in autopilot to make their ratings and questionnaire responses.

MEDIATIONAL EXPERIMENT RESULTS

The main results of the mediational experiment are presented in Table 2. The table includes individual ANOVA's at a corrected α level of 0.01, standardized mean (Z-score) values for each rating scale, and Duncan's multiple comparisons tests for those scales having significant ANOVA's.

The results indicate that only the FBMCH scale was not significant at $p < 0.01$. All of the scales exhibited monotonic increases with load. In terms of the Duncan's tests, sensitivity among those scales demonstrating significance could be ranked as follows: Most sensitive, MCH+; next most sensitive, COMPMCH and NDT; next most sensitive, MCH and COMPMCH+. However, all five scales are actually quite sensitive, considering the small sample size and strict criterion used.

To provide substantiation of the results obtained with the rating scale data, a MANOVA was performed using both mean response time and percentage of errors on the navigation problems for each experimental flight as dependent measures. When using the F-approximation of Wilks U-statistic to compare the groups of subjects assigned to each rating scale condition, there was no significant main effect of rating scale, $F(10,58) = 1.49$, $p = 0.1684$. This result indicates that no differences in primary task performance were associated with subject assignment. The lack of a rating scale main effect suggests that conclusions regarding the sensitivity of the scales are based on true scale differences rather than group differences in primary task performance.

A second MANOVA was conducted to determine whether there was a main effect of experience level on mean response time and percent error in the mediational task. The F-approximation to Wilk's U-statistic revealed no significant differences in task performance associated with experience level, $F(10, 58) = 0.49$, $p = 0.8894$.

Using the standardized ratings for the three load presentations--first, second, or third, a one-way analysis of variance revealed no significant differences attributed to load level presentation order, F (2,70) = 0.37, p = 0.6942. A one-way ANOVA on the sum of the standardized ratings across the load levels for each subject indicated no significant effects of experience level on the summed ratings, F (15,30) = 1.33, p = 0.2815.

The questionnaire responses to the low, medium, and high load levels were sorted into comments which were "positive" or favorable in tone and "negative" or unfavorable in tone. A Chi-square test revealed significant differences in the frequencies of the favorable and unfavorable responses across the load levels, χ^2_2 = 55.94, p = 0.0001. Favorable comments occurred most often at the low load level, while unfavorable ones occurred most often at the high load level. Based on categories which were derived by sorting, it seems that the major factors which influenced the subjects' ratings were the amount of time available, the difficulty of the task, and their assessment of how well the task requirements were met.

In terms of comparison of the MCH scale results of this experiment with those of the earlier mediational experiment (Wierwille, Rahimi, and Casali, 1984), it was found that again the two were virtually identical.

CONCLUSIONS DRAWN FROM THE RESULTS OF THE TWO EXPERIMENTS

Several conclusions can be readily drawn by comparing the information contained in Tables 1 and 2. First, in terms of global sensitivity, only the MCH and the COMPMCH exhibited significance in both experiments at the $p < 0.01$ level. This finding indicates that none of the other scales possess as high a general sensitivity as the MCH scale and its computerized version. All of the other scales exhibited sensitivity in only one experiment. While the MCH+ scale and NDT scale exhibited slightly higher sensitivities than the MCH in the mediational experiment, these two scales could not be counted on to provide better results than the MCH in other types of experiments.

The table also shows that the MCH scale and COMPMCH scale are about equal in sensitivity. Apparently, computerizing the scale, such that a subject is forced to use the tree structure, has no effect on the sensitivity of the scale. In the communications experiment the MCH scale is slightly more sensitive, and in the mediational experiment the COMPMCH is slightly more sensitive. On balance, however, they have the same sensitivity.

It should be noted that each given subject used only one rating scale. Thus, the ratings for the MCH+ scale, for example, were performed by the same group of subjects in both experiments. Therefore, one cannot attribute the differences in scale sensitivity across experiments to individual differences in subject groups. All other peripheral statistical tests support the conclusion that all of the scales except the MCH and COMPMCH are task dependent.

Other conclusions can also be drawn. Does increasing the number of categories from 10 to 15 as in the MCH+ scale (Figure 2) improve sensitivity? The answer appears to be "not consistently". While the MCH+ is somewhat more sensitive in the mediational experiment, it is substantially less sensitive in the communications experiment. For the computerized version of the 15 category scale (the COMPMCH+), sensitivity is about the same as the MCH in the mediational experiment and much lower than the MCH in the communications experiment. The conclusion is that 15 categories is not generally as good as 10 categories.

Does revision of the scale to produce a left-to-right decision tree with 15 categories (the PBMCH, Figure 3) improve sensitivity? The answer to this question is "no". The PBMCH is not as sensitive as the MCH in either of the two experiments.

Finally, does a tabular format, with the decision tree removed (the NDT, Figure 4) improve sensitivity? The answer in this case is again "not consistently". While the NDT is slightly more sensitive than the MCH in the mediational experiment, it is much less sensitive than the MCH in the communications experiment.

In regard to the questionnaire responses, it was found that pilots do rate on the basis of concepts similar to those which researchers tend to think should be included in workload. While wording did vary, the subjects tended to rate on the basis of time pressure, difficulty, assessed performance, and problems of time sharing. Their comments changed in tone and frequency as expected with load level.

In general then, conflicting results between the two experiments indicate that sensitivity of most rating scales varies in subtle ways. However, the MCH scale and its computerized version are consistently sensitive and reliable. Furthermore, pilots' ratings appear to be based on factors similar to those which researchers currently consider important.

ACKNOWLEDGEMENTS

The authors wish to thank Ms. Sandra Hart, NASA Ames Research Center, who served as grant technical monitor for this project (NAG2-17). The authors also wish to thank Drs. John G. Casali, Robert D. Dryden, and Harry L. Snyder for their helpful suggestions.

REFERENCES

Casali, J. G. and Wierwille, W. W. A comparison of rating scale, secondary task, physiological, and primary task workload estimation techniques in a simulated flight task emphasizing communications load. Human Factors, 25 (6), December, 1983, 623-641.

Cooper, G. E. and Harper, R. P., Jr. The use of pilot rating in the evaluation of aircraft handling qualities. Moffett Field, CA: National Aeronautics and Space Administration, Ames Research Center, NASA TN-D-5153, April, 1969.

North, R. A. and Graffunder, K. Evaluation of a pilot workload metric for simulated VTOL landing tasks. Proceedings of the Human Factors Society Twenty-Third Annual Meeting, Boston, Massachusetts, 1979, 357-361.

O'Conner, M. F. and Buede, D. M. The application of decision analytic techniques to the test and evaluation phase of the acquisition of a major air system. McLean, Virginia: Decisions and Designs, Inc., Report TR 77-3, 1977.

Seifert, R., Daniels, A. F., and Schmidt, K. A method of man-display/control system evaluation. Proceedings of the AGARD Conference on Guidance and Control Displays, No. 96, AGARD-CP-96, February, 1972, 8-1--8-9.

Wierwille, W. W. and Connor, S. A. Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. *Human Factors*, *25* (1), February, 1983, 1-16.

Wierwille, W. W. and Casali, J. G. A validated rating scale for global mental workload measurement applications. *Proceedings, 27th Annual Meeting of the Human Factors Society*, Norfolk, Virginia, October, 1983, Vol. 1, 129-133.

Wierwille, W. W., Rahimi, M., and Casali, J. G. Evaluation of sixteen measures of mental workload using a simulated flight task emphasizing mediational behavior. *Human Factors*, *25*, 1984. (In press, special issue, Human Factors in Aviation Psychology.)

Wolfe, J. D. Crew workload assessment: Development of a measure of operator workload. Minneapolis, Minnesota: Honeywell, Inc. Technical Report No. AFFDL-TR-78, 1978.

TABLE 1. Communications Experiment Results

Rating Scale	ANOVA F (2, 10) p	Standardized Mean Scores* L	M	H
MCH	0.0035	-1.011	0.395	0.616
COMPMCH	0.0015	-0.510	0.000	0.510
MCH+	0.1347	-0.594	-0.019	0.613
COMPMCH+	0.0433	-0.484	0.132	0.352
PBMCH	0.0012	-0.983	0.046	0.937
NDT	0.0237	-0.522	-0.095	0.617

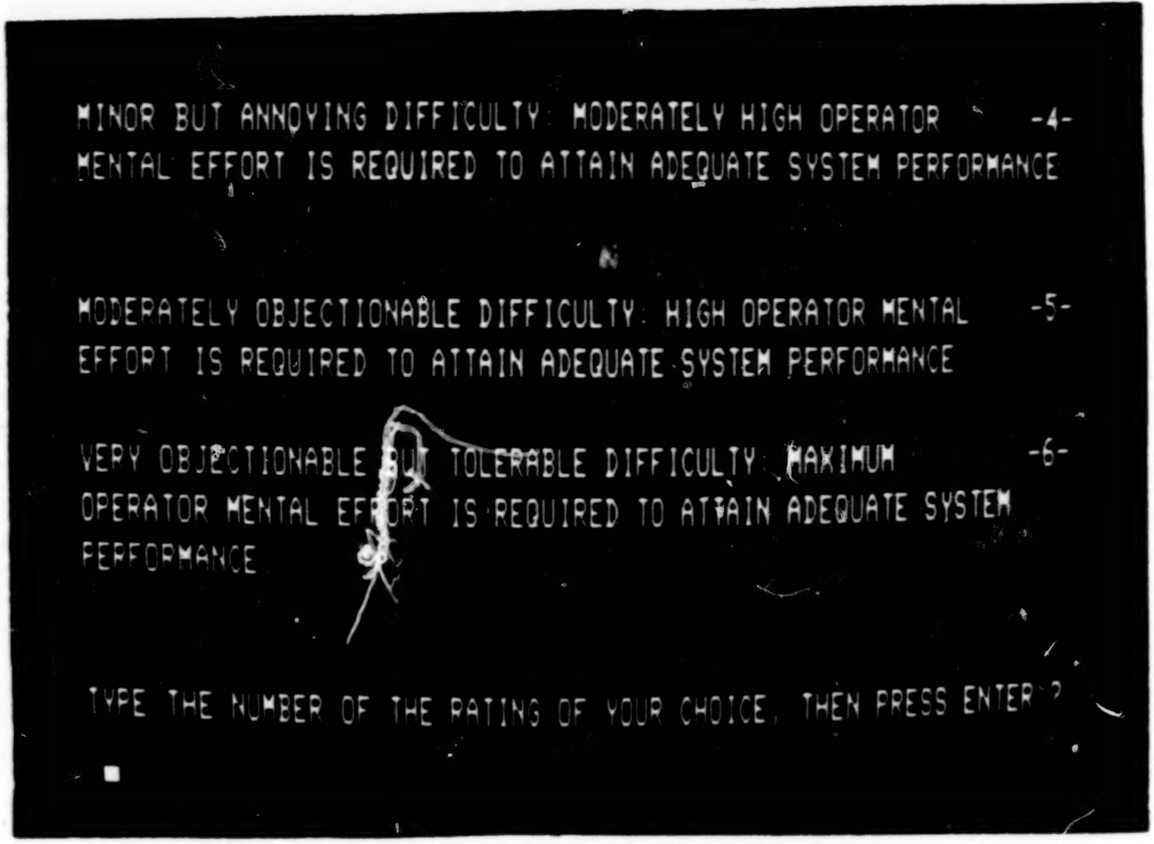
*Means with a common underline do not differ at $p < 0.01$ using Duncan's multiple comparisons. These comparisons were performed only on rating scale scores demonstrating significance at $p < 0.01$ in the individual ANOVAs.

TABLE 2. Mediational Experiment Results

Rating Scale	ANOVA F (2, 10) p	Standardized Mean Scores* L	M	H
MCH	0.0039	-0.758	-0.188	0.946
COMPMCH	0.0030	-0.667	-0.200	0.867
MCH+	0.0001	-1.023	-0.440	1.068
COMPMCH+	0.0064	-0.808	-0.195	0.613
PBMCH	0.3800	-0.279	-0.082	0.361
NDT	0.0001	-0.845	-0.307	1.152

*Means with a common underline do not differ at $p < 0.01$ using Duncan's multiple comparisons. These comparisons were performed only on rating scale scores demonstrating significance at $p < 0.01$ in the individual ANOVAs.

ORIGINAL PAGE IS
OF POOR QUALITY



MINOR BUT ANNOYING DIFFICULTY: MODERATELY HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE -4-

MODERATELY OBJECTIONABLE DIFFICULTY: HIGH OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE -5-

VERY OBJECTIONABLE BUT TOLERABLE DIFFICULTY: MAXIMUM OPERATOR MENTAL EFFORT IS REQUIRED TO ATTAIN ADEQUATE SYSTEM PERFORMANCE -6-

TYPE THE NUMBER OF THE RATING OF YOUR CHOICE, THEN PRESS ENTER ?

■

Figure 1. Sample frame of the COMPMCH rating scale.

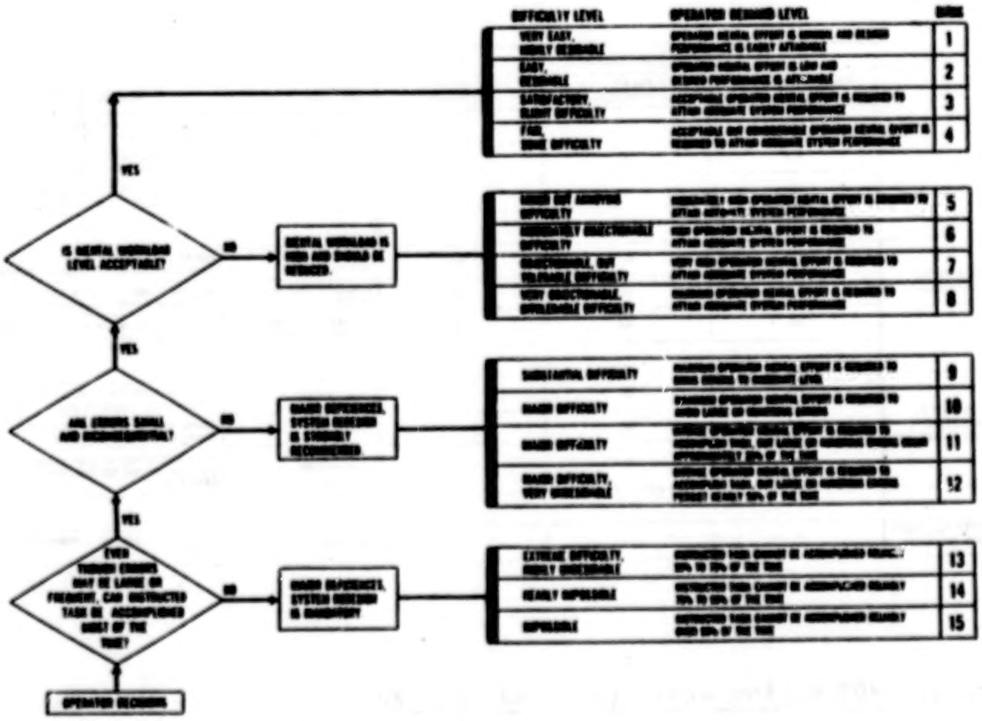

DIFFICULTY LEVEL
OPERATOR DEMAND LEVEL
VERY EASY, HIGHLY DESIRABLE
1
EASY, DESIRABLE
2
SATISFACTORY, SLIGHT DIFFICULTY
3
FAIR, SOME DIFFICULTY
4
5
6
7
8
SUBSTANTIAL DIFFICULTY
9
MAJOR DIFFICULTY
10
MAJOR DIFFICULTY
11
MAJOR DIFFICULTY, VERY UNDESIRABLE
12
EXTREME DIFFICULTY, HIGHLY UNDESIRABLE
13
NEARLY IMPOSSIBLE
14
IMPOSSIBLE
15
YES
IS MENTAL WORKLOAD LEVEL ACCEPTABLE?
NO
YES
ARE ERRORS SMALL AND INCONSEQUENTIAL?
NO
YES
NO
OPERATOR DECISIONS

Figure 2. MCH+ rating scale (reduced in size).

C-2

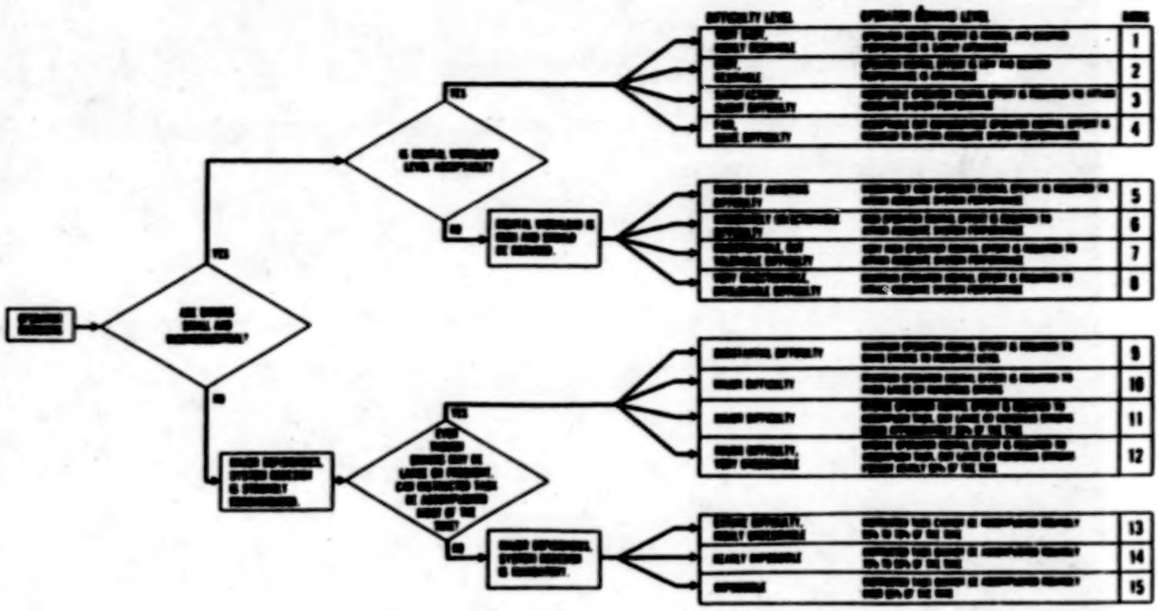

DIFFICULTY LEVEL
OPERATOR DEMAND LEVEL
YES
NO
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Figure 3. PBMCH rating scale (reduced in size).

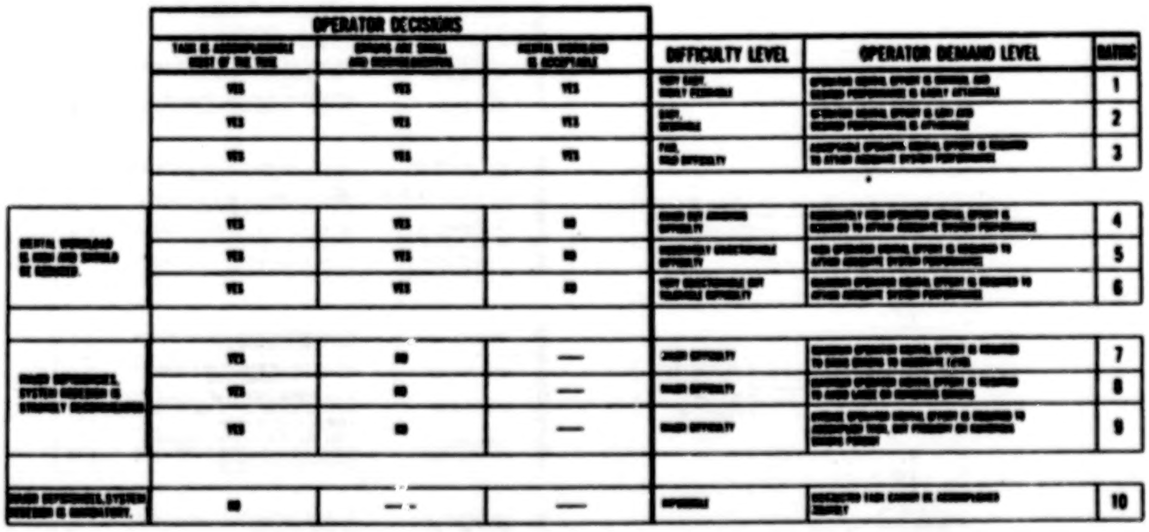

OPERATOR DECISIONS
DIFFICULTY LEVEL
OPERATOR DEMAND LEVEL
1
2
3
4
5
6
7
8
9
10

Figure 4. NDT rating scale (reduced in size).

ORIGINAL PAGE IS
OF POOR QUALITY

N85
14542

UNCLAS

N85 14542

D7

ASSESSING THE SUBJECTIVE WORKLOAD OF DIRECTIONAL ORIENTATION TASKS

Ronald C. Miller
Informatics General Corporation (PSOW)
Palo Alto, CA

Sandra G. Hart
NASA-Ames Research Center
Moffett Field, CA

ABSTRACT

An experiment was conducted to investigate the impact of various flight-related tasks on the workload imposed by the requirement to compute new headings, course changes and reciprocal headings. Eight instrument-rated pilots were presented with a series of heading-change tasks in a laboratory setting. Two levels of difficulty of each of three tasks were presented verbally (numeric values imbedded in simple commands) and spatially (headings were depicted on a graphically drawn compass). Performance was measured by evaluating the speed (response times) and accuracy (percent correct and time outs) of the responses. The workload experienced by the pilots under each experimental condition was determined by responses to a standard set of bipolar rating scales. The subjective responses and objective measures of performance reflected a strong association between subjective experience and objective behavior. The reciprocal calculations were performed quickly and accurately throughout and were considered to be minimally loading. Subjective workload, percent correct and response times for the two course-change tasks varied significantly as a function of level of difficulty and display format, with no discernable speed/accuracy trade off. The results of this study will be used to predict the workload that is imposed on pilots of actual and simulated flights by course corrections and computations in conjunction with previously obtained estimates of control and communications workload.

INTRODUCTION

This is one of a series of experiments designed to investigate the relationships among the demands that are imposed on pilots, the levels of workload they experience, and their performance of flight-related tasks. Flight tasks that impose objectively and independently determined levels of workload on pilots must be identified to assist in the creation of simulation scenarios. Such predictive validity is central to the successful conduct of simulation research so as to avoid the circularity involved in ad hoc scenario creation coupled with post-hoc evaluation. All too often, the success of a load manipulation is determined by referencing the same measures of workload or performance that the simulation was designed to investigate in the first place.

In two previous studies (Childress, Hart & Bortolussi, 1982; Hart & Bortolussi, 1984), pilots were asked to give their opinions about the potential impact of a number of routine, as well as unusual, flight-related tasks on their subsequent performance, workload, stress, and effort. In order to evaluate the accuracy of the opinion measures, a series of simulations was conducted (Kantowitz, Hart & Bortolussi, 1983; Kantowitz, Hart, Bortolussi, Shively & Kantowitz, 1984) to determine the actual impact of a subset of the discrete events (e.g., communications, flight-plan changes) and routine, continuous control activities (e.g. speed, heading, and altitude control) on pilots' subjective experiences, performance, and ability to accomplish an additional concurrent task.

In the current experiment, the influence of specific navigation-related tasks on pilot workload and performance was investigated to expand the breadth of flight-related tasks for which levels of load could be objectively and independently predicted. This study imposed three types of mental arithmetic calculations that are encountered inflight: calculating reciprocal headings, determining new headings and calculating the number of degrees turned to a new heading. By imposing relatively realistic tasks in the laboratory, the number of computational steps and the type of processing (verbal or spatial) performed by the subjects could be varied to experimentally impose different levels and types of workload. In addition, a large number of responses could be obtained in a relatively short period of time. Upon completion of each task, subjective workload ratings were elicited and used to determine the relative subjective workload differences between tasks and types of processing within tasks.

In earlier laboratory research (Loftus, 1978), performance on various spatial direction change tasks was investigated to determine the effect of direction of change (clockwise or counter-clockwise) on computational speed and accuracy. It was found that the direction of turn made no significant difference in the difficulty of the tasks. Thus, the "EASY" and "HARD" manipulations used in the current study were determined by whether or not the turn passed through 360 degrees, rather than by the direction of turn. An equal number of Left and Right turns were presented within each block of trials.

The navigation problems were presented spatially and verbally to investigate the effect of display format on computational speed and accuracy and the associated variation in experienced workload. With the introduction of computers into the cockpits of aircraft, analog dials and compasses are being replaced by alphanumeric and stylized graphic displays. Thus, the mental workload associated with computing (or monitoring changes in) desired or current heading or ground path may be considerably different if the information is presented on a traditional compass rose or as discrete alphanumeric values. With traditional instruments, no computations may be involved at all. Pilots may simply observe the current heading, find the new heading on the indicator and maneuver the aircraft in the appropriate direction until the desired heading has been achieved without ever calculating the amount of change or the exact value of the new heading. This may not be the case with computer displays of navigation information. New headings or course changes may be entered by keyboard into an onboard computer as discrete numerical values, rather than as rough spatial locations. Furthermore, monitoring the rate of change or estimating the amount of change yet to occur from a digital display may require different mental processes than judging angular deviations from a dial.

For this reason, current headings were displayed in two different ways in the current study: graphically on a stylized compass rose (COMPASS), and alphanumerically (ALPHA). The new heading (or amount of required change) was always presented in a text format, simulating a verbal command from an Air Traffic Controller to adopt a new course (e.g. "TURN RIGHT 90 DEGREES" or "FLY HEADING 090, etc.)

METHOD

Subjects

One female and seven male instrument-rated pilots, ranging in experience from a newly rated pilot with 45 hours instrument time to an Air Transport pilot with more than 5000 hours of instrument flying participated in the study. The pilots ranged in age from 20 to 45 years.

Apparatus

The tasks were presented on a 22.9 cm computer monitor driven by an APPLE II+ computer. Responses were entered via a numeric keyboard. Subjects were seated 55 cm in front of the display, which was located in a small, dimly lighted response booth . Response times were recorded to the nearest 10 msecs. If the subject failed to enter a response within 20 sec from the onset of the trial, the display disappeared and a "time out" was recorded.

Task Description

Reciprocal

Subjects were given a current heading, (e.g. 005 degrees) and asked to enter the reciprocal heading, (in this case 185 degrees). (Fig. 1) The RECIPROCAL tasks were presented with two levels of difficulty. In the EASY trials, the initial heading ranged from 0-180 degrees. On the HARD trials the initial heading ranged from 181 to 359 degrees. It was predicted that adding 180 degrees to the current heading to obtain the reciprocal heading would prove to be easier than subtracting 180 degrees.

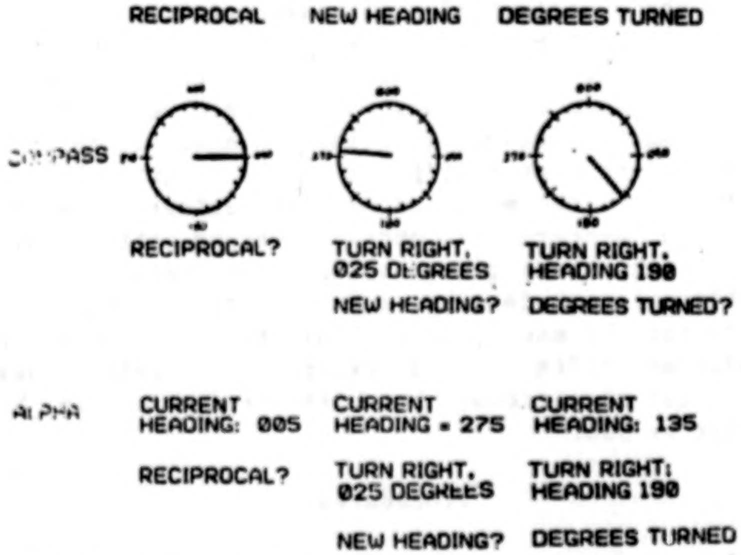

RECIPROCAL
NEW HEADING
DEGREES TURNED
COMPASS
270
180
RECIPROCAL?
TURN RIGHT, 025 DEGREES
NEW HEADING?
TURN RIGHT, HEADING 190
DEGREES TURNED?
ALPHA
CURRENT HEADING: 005
RECIPROCAL?
CURRENT HEADING = 275
TURN RIGHT, 025 DEGREES
NEW HEADING?
CURRENT HEADING: 135
TURN RIGHT: HEADING 190
DEGREES TURNED

Figure 1: Display formats.

New Heading

A current heading was presented (e.g. 275 DEGREES) followed by a direction (LEFT or RIGHT) and the number of degrees to turn (e.g. 25 DEGREES). The subject entered the new heading (in this case, 300 degrees). (Fig. 1) Once again, two levels of difficulty were presented. The EASY condition consisted of a heading change that did not pass through 360 degrees, while the HARD condition consisted of a heading change that did pass through 360 degrees.

Degrees Turned

In the Degrees Turned task, subjects were presented once again with a current heading (e.g. 135 DEGREES) and a direction to turn (e.g. RIGHT). They were then given a new heading (e.g. 190 DEGREES) and asked to respond with the number of degrees of turn required to achieve the new heading (in this case, 55 DEGREES). (Fig. 1) As with the New Heading task, the EASY version did not require a turn through 360 degrees while the HARD condition did require a turn through 360.

Bipolar Rating Scales

The pilots completed 10 rating scales at the conclusion of each task. The scales represented dimensions of the task (e.g. difficulty, time pressure), operator-related variables (e.g. fatigue, stress), types of effort expended (e.g. physical, mental/sensory), the pilots assessment of their own performance and overall workload. (Kantowitz, Hart & Bortolussi, 1983; Kantowitz, Hart, Bortolussi, Shively & Kantowitz, 1984; Hart, Battiste & Lester, 1984) The computer-generated scale, consisted of an 11-cm. vertical line bounded by relevent descriptors (e.g. Extremely Easy / Impossible) presented sequentially. The ratings were entered by positioning a joystick-controlled cursor to a point on the line corresponding to the subject's perception of the magnitude of the displayed element. The subjects ratings were quantified by assigning numerical values from 1-100 to the length of the scale.

Immediately before the experimental trials were given, the pilots were asked to evaluate the subjective importance of each of the nine dimensions to their own definition of overall workload (Hart, Battiste & Lester, 1984). The importance attached to each factor was entered as a weight (from a possible low value of 0 to a maximum value of 8) into an equation in which the bipolar ratings for each subject were multiplied by the weights obtained from each subject for that factor. The weighted bipolar ratings for each experimental condition, were then averaged to obtain a second estimate of the overall workload of each task, taking into account the relative contribution of different, possibly relevant factors, and the different subjective importance placed on the factors by each pilot. This procedure has been found to maintain the relative differences in experienced workload associated with different experimental conditions, but with a reduction in between-subject variability as great as 50%.

Procedures

Three directional orientation tasks, RECIPROCAL, NEW HEADING and DEGREES TURNED were presented to each subject in blocks of 20 trials. Only one type of task was presented within each block. Each task was presented with each of two levels of difficulty, in both a spatial (COMPASS) and verbal (ALPHA) format. There was however, no mixing of display types or difficulty levels within a given block. (Fig. 2) All initial headings, amounts of change and new headings were given with a least-significant digit of 0 or 5 to reduce interpretation errors in the COMPASS conditions.

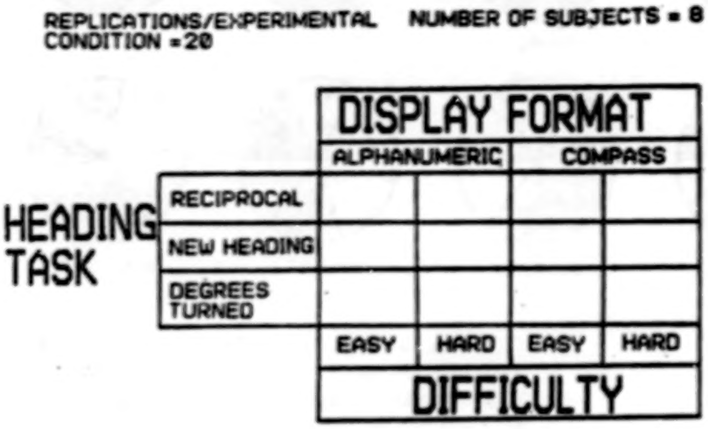


Figure 2: Experimental design.

The twelve experimental conditions were presented in a different random order to each pilot. Following general instructions about the purpose of the experiment, the pilots were familiarized with the apparatus and each of the experimental tasks was described and demonstrated. Five practice trials (with performance feedback) were given prior to the presentation of each experimental task. The pilots were given no feedback about their response times or accuracy during the experimental trials. Immediately after each block of 20 experimental trials, the pilots were asked to respond to 10 bipolar rating scales to evaluate their experiences during the immediately preceeding experimental conditions. The entire experiment lasted about 3 hr.

RESULTS AND DISCUSSION

The performance measures that were collected for analysis included response times, number of correct responses and number of time outs. The trials that ended in time outs were counted as incorrect responses, and were not included in calculating mean response times.

Three-way analyses of variance for repeated measures were conducted on the three measures of performance and the rated workload. In addition, individual two-way analyses of variance for repeated measures were performed on each task individually.

Rating Scales

As has been found in a number of previous experiments (Kantowitz, Hart & Bortolussi, 1983; Kantowitz, Hart, Bortolussi, Shively & Kantowitz, 1984; Hart, Battiste & Lester, 1984), there were large differences between subjects in the relative importance each placed on the nine dimensions. (Fig. 3) Time

ORIGINAL PAGE IS
OF POOR QUALITY

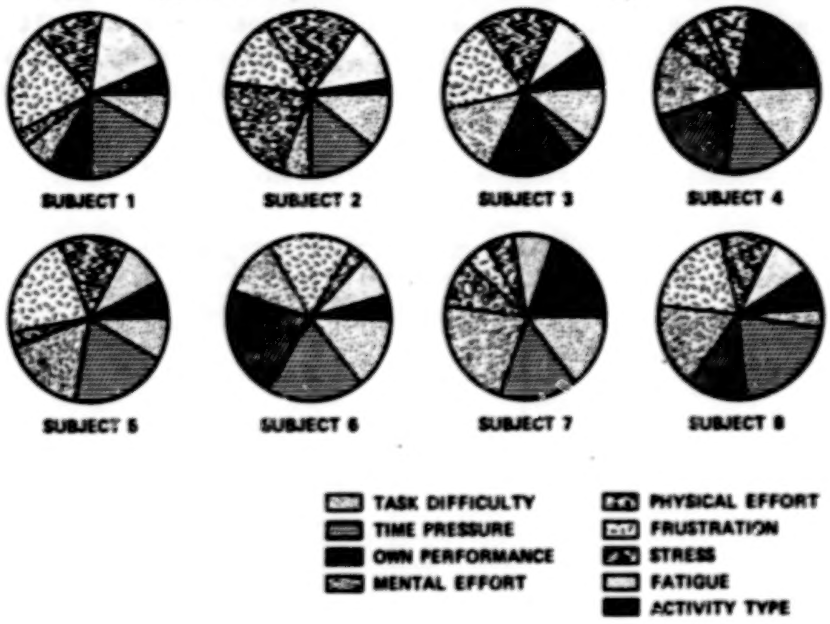


Figure 3: Relative importance to the subjective experience of workload of each each of nine factors. (n = 8)

Pressure, Mental/Sensory Effort, Stress and Frustration were generally considered to play an important role in generating different levels of experienced workload. There was considerable disagreement about the influence of Performance and Physical Effort although Physical Effort was considered to be a relatively unimportant factor by most of the pilots.

Table 1. Average bipolar ratings. (n = 8)

	Rated workload-related dimension									
	Task Diff.	Time Press.	Performance	Ment. Eff.	Phys. Eff.	Frustration	Stress	Fatigue	Act. Lvl.	Ov'all Wkld.
C/Recp/E	23	27	26	36	11	25	26	19	42	22
C/Recp/H	22	26	23	30	13	25	29	26	42	23
C/NewH/E	54	44	49	55	14	49	38	36	53	48
C/NewH/H	62	58	59	58	14	58	40	34	57	59
C/DegT/E	39	41	46	46	14	44	37	28	56	47
C/DegT/H	56	53	56	57	14	49	39	30	57	58
A/Recp/E	28	23	27	40	13	31	23	23	49	25
A/Recp/H	25	31	30	36	11	30	26	22	43	28
A/NewH/E	39	31	41	42	10	34	26	28	49	39
A/NewH/H	63	60	62	62	22	60	42	34	58	64
A/DegT/E	22	26	24	31	14	21	20	22	47	24
A/DegT/H	40	40	50	50	14	38	28	27	53	45

The bipolar ratings obtained may be seen in Table 1. The raw bipolar ratings were individually weighted (to account for individual subjects' biases about which task dimensions most affected their experience of workload) and averaged. This weighting procedure reduced the overall between-subject standard deviation from 21 to 9.5 relative to that obtained with the single global rating of overall workload. This reduction in between-subject variability occurred for each experimental condition, individually, as well as across experimental conditions.

In general, the workload imposed by the experimental tasks was considered to be moderate. The RECIPROCAL task (mean rating = 37) was rated as significantly less loading (F(2,14) = 16.28, p<0.01) than the NEW HEADING (mean rating = 50) or DEGREES TURNED (mean rating = 44) tasks. There was a significant difference (F(1,7) = 6.38, p<0.05) in rated workload as a function of display condition (COMPASS Rating = 46, ALPHA Rating = 44), and Difficulty Level (F(1,7) = 11.34, p<0.05), (EASY rating = 41, HARD rating = 45). In addition interesting differences were found on a task-by-task basis.

Comparisons Within Tasks

Reciprocal

In general, the RECIPROCAL task was performed accurately (average percent correct = 90) and quickly (average response time = 6.73 sec). There were few time outs and workload was considered to be low. (Fig. 4) There were no significant differences in the percent correct or mean response times for the RECIPROCAL task between either the EASY/HARD conditions or the COMPASS/ALPHA display formats,nor were there any significant DIFFICULTY by DISPLAY FORMAT interactions. There were no significant differences between workload ratings between the EASY/HARD or COMPASS/ALPHA conditions, nor were there significant differences in the number of time outs.

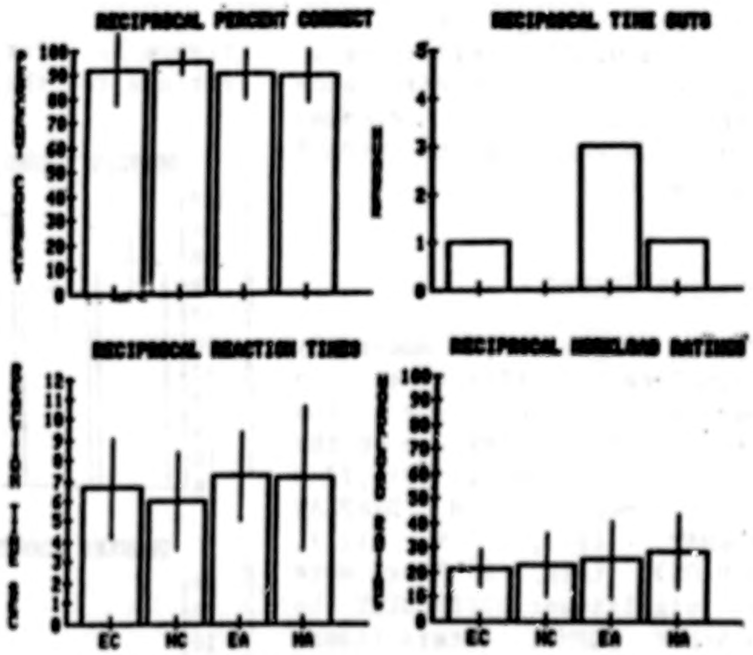


Figure 4: Summary of performance data and workload ratings for the RECIPROCAL task.

New Heading

For the NEW HEADING task, there was a significant, (F(1,7) = 22.89, p<0.01) decrease in the percent of correct responses between the EASY/HARD conditions and a significant increase in response times (F(1,7) = 53.98, p<0.001). (Fig. 5) There were no significant differences in percent correct or response times between the EASY/HARD or COMPASS/ALPHA conditions, nor was there a significant

DIFFICULTY by DISPLAY FORMAT interaction. There was a significant difference in the number of time outs between the EASY/HARD conditions ($F(1,7) = 13.49$, $p<0.01$). There were 4.5 time outs (out of 20 possible trials) with the HARD condition, and only 1.5 with the EASY condition. There was no significant difference in time outs between COMPASS/ALPHA. A significant increase in the weighted overall workload ratings (from 46 to 54) was found between the EASY and HARD conditions ($F(1,7) = 13.93$, $p<0.01$) but no differences were found as a function of DISPLAY FORMAT. There was a significant interaction between EASY/HARD and DISPLAY FORMAT for workload ratings ($F(1,7) = 8.13$, $p<0.05$), reflecting a greater EASY/HARD difference for the ALPHA display condition than for the COMPASS display condition.

Degrees Turned

There were moderatly significant differences in percent correct for the DEGREES TURNED task due to the level of difficulty ($F(1,7) = 11.9$, $p<0.05$) and DISPLAY FORMAT ($F(1,7) = 11.7$, $p<0.05$). (Fig. 6) There were no significant DIFFICULTY by DISPLAY FORMAT interactions. For response time, there were highly significant differences due to the EASY/HARD manipulation ($F(1,7) = 31.3$, $p<0.01$) and the COMPASS/ALPHA formats ($F(1,7)=100.7$, $p<0.001$). There was a significant DIFFICULTY by DISPLAY FORMAT interaction ($F(1,7) = 18.65$, $p<0.01$). An examination of the mean response latencies for the

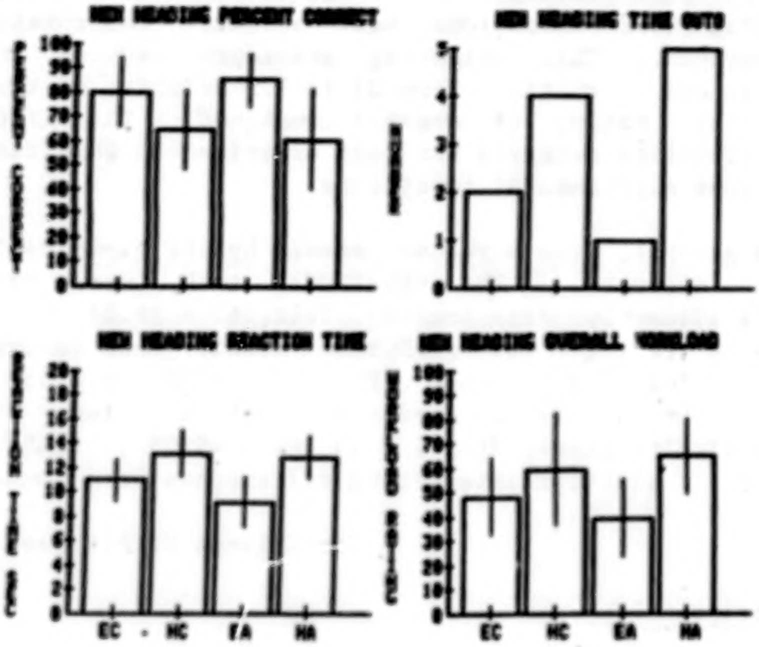


Figure 5: Performance data and workload ratings for the NEW HEADING tasks. (n=8)

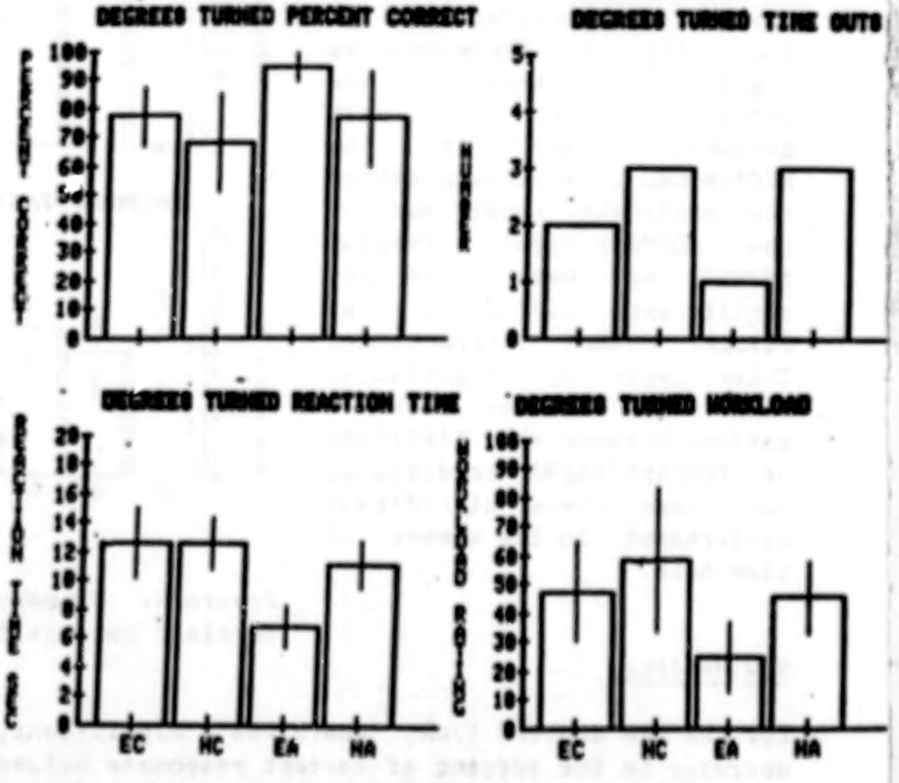


Figure 6: Performance data and workload ratings for the DEGREES TURNED tasks.

EASY/HARD conditions shows an increase from 9.73 to 11.86 sec between the EASY and HARD conditions. The mean response times were 3.66 sec greater for the COMPASS display format than for the ALPHA display format, suggesting overall differences between both levels of difficulty and display types. The significant interaction suggests otherwise, however. The means for EASY COMPASS and HARD COMPASS are 12.61 and 12.62 sec respectively; the effect of varying the difficulty level is non-existent. For EASY ALPHA and HARD ALPHA however, the result is different. The EASY ALPHA calculations were performed very quickly (6.18 sec), while HARD ALPHA calculations took 11.01 sec. Thus, the effect of varying the level of difficulty had a major impact on response time in the ALPHA condition but not the COMPASS condition.

There were no significant differences in the number of time outs between EASY/HARD and COMPASS/ALPHA and no significant differences in the overall workload ratings between EASY and HARD. There was a significant increase in the weighted overall workload ratings (F(1,7) = 37.22, p<0.001), between the ALPHA presentation format (mean rating = 39) and the COMPASS presentation format (mean rating = 48).

Influence of Display Format

An examination of the COMPASS/ALPHA conditions collapsed across tasks and difficulty levels produced some interesting results. There were no sigificant differences between EASY/HARD for either percent correct or response time for the COMPASS condition despite the fact that EASY/HARD differences showed up in two of the three tasks. In the ALPHA condition, there were significant EASY/HARD differences for percent correct (F(1,7) = 18.77, p<0.01) and response time (F(1,7) = 37.59, p<0.001), which would lead to the conclusion that the greater portion of the EASY/HARD variance can be attributed to the ALPHA condition. The between-tasks differences were significant for the COMPASS condition for both percent correct (F(1,7) = 20.9, p<0.01) and response time (F(1,7) = 68.6, p<0.001). There was a significant between-tasks difference, as well, for percent correct (F(1,7) = 11.63, p<.05) and response time (F(1,7) = 21.09 p<0.01). Since there were no EASY/HARD or COMPASS/ALPHA differences for the RECIPROCAL task, and performance on this task was considerably different than on the other two tasks, a three-way analysis of variance for repeated measures was performed on the performance measures with the RECIPROCAL data omitted. For the COMPASS condition, EASY/HARD differences were found for percent correct (F(1,7) = 7.3, p<0.05) that were not evident with RECIPROCAL included, although differences in response time did not acheive significance. No significant difference was found between the two heading tasks. For the ALPHA conditions, a significant decrease in percent correct was still found (F(1,7) = 25.86, p<0.01) as was a significant increase in response latency (F(1,7) = 80.96, p<0.001) between the EASY and HARD conditions. In addition, a significant difference was found between the two heading tasks for percent correct (F(1,7) = 7.6, p<0.05) and response latency (F(1,7) = 20.15, p<0.01).

Comparison Between Tasks

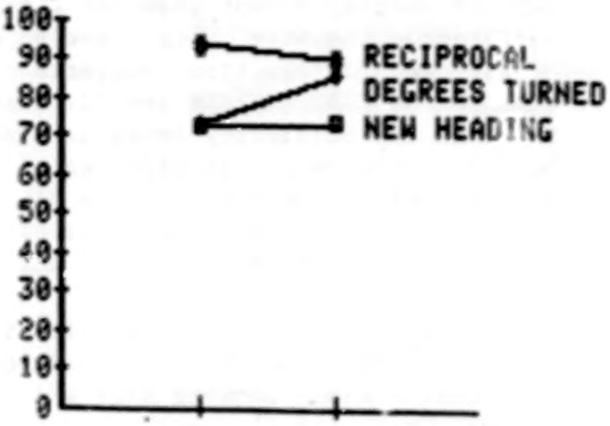
The results suggest that the calculation of reciprocals is a task in a class by itself. Neither the differences in initial headings or display format made any difference in the percent correct, response times or workload ratings. (Fig. 7) Overall, the percent correct was greater, the response times faster and workload ratings lower than for either of the other two tasks.

This is probably due to the use of rules-of-thumb to calculate reciprocals. The most commonly used rule-of-thumb requires the pilot to add 200 and subtract 20 if the initial heading is between 0 and 180 or to subtract 200 and add 20 if the initial heading is between 180 and 360. Although this rule may result in answers that pass through 360 degrees the additional calculation required to provide a response between 0 and 360 is relatively trivial. This implies that the calculation of reciprocal headings would not be a particularly useful task to generate variations in workload, although it does provide a reliable, albeit low, amount of workload.

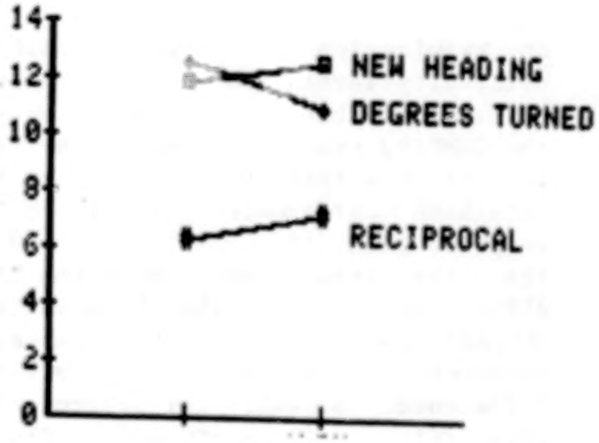
For the NEW HEADING task, in general, the response times for the EASY conditions were lower, the number of time outs less, the percent correct higher and workload ratings lower than for the HARD conditions. The DEGREES TURNED task produced a highly significant difference between the EASY and HARD conditions for response times. The EASY and HARD conditions were relatively equal as far as the number of time outs incurred. Percent correct was higher for the EASY condition than HARD. It was clear that there was no speed accuracy trade off in any of the experimental conditions.

For the COMPASS tasks in general, the response times and number of time outs were approximately equal between EASY and HARD and the percent correct was somewhat higher for the EASY condition. Across tasks, in the COMPASS condition (with the RECIPROCAL task removed), there were no significant differences between NEW HEADING and DEGREES TURNED. For the ALPHA condition, however, there was a significant difference between NEW HEADING and DEGREES TURNED (for NEW HEADING there were fewer correct responses, higher response times and a greater number of time outs than for DEGREES TURNED). In addition, the impact of the EASY/HARD manipulation was greater with the ALPHA display than with the COMPASS display.

a. Percent correct.



b: Response latency (sec)



c. Workload rating

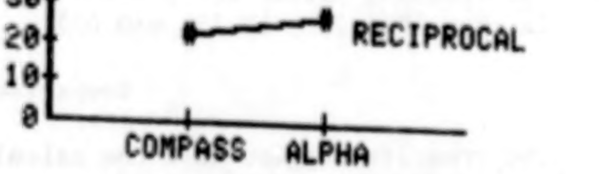


Figure 7: Interactions among tasks and display types. (n = 8)

CONCLUSIONS

The results of this study suggest that the NEW HEADING task is more sensitive than the the DEGREES TURNED task to changes in difficulty levels in the areas of correct responses and overall workload ratings. The DEGREES TURNED task seems to be more capable of generating response time differences.

In general, the COMPASS format did not generate differences due to task type or difficulty. The use of the compass by instrument-rated pilots as an aid in solving heading-related problems, seems to preclude the use of heading-change tasks to produce performance and workload differences when a compass is available. The sensitivity of the ALPHA format however, suggests that heading-related computations can be used, in some instances, to induce performance and workload differences. Replication of these results under more realistic inflight and simulation conditions is required to establish the validity and reliability of these results for imposing predictable levels of load and performance.

REFERENCES

Childress, M. E., Hart, S. G. & Bortolussi, M. R. The Reliability and Validity of Flight Task Workload Ratings. Proceedings of the Human Factors Society-26th Annual Meeting, Seattle Wa. -1982, 319-323

Hart. S. G., Battiste, V. & Lester, P. POPCORN: A complex Supervisory Control Simulation for Workload and Performance Research. Proceedings of the Twentieth Annual Conference on Manual Control., Moffett Field, Ca. -1984, (in press)

Hart, S. G. & Bortolussi, M. R. Pilot Errors as a Source of Workload. Human Factors , 1984, (in press)

Kantowitz, B. H., Hart, S. G. & Bortolussi, M. R. Measuring Pilot Workload in a Moving Base Simulator: I. Asynchronous Secondary Choice-Reaction Task. Proceedings of the Human Factors Society-27th Annual Meeting , Norfolk, Va.-1983, 319-323

Kantowitz, B. H., Hart, S. G., Bortolussi, M. R., Shively, R. J. & Kantowitz, S. C. Measuring Pilot Workload in a Moving Base Simulator: II. Building Levels of Workload. Proceedings of the Twentieth Annual Conference on Manual Control. , Moffett Field, Ca.-1984, (in press)

Loftus, G. R. Comprehending Compass Directions. Memory and Cognition 1978, Vol.6 (4), 416-422

N85
14543

UNCLAS

D8

CLASSIFICATION SYSTEMS FOR INDIVIDUAL DIFFERENCES IN MULTIPLE-TASK PERFORMANCE AND SUBJECTIVE ESTIMATES OF WORKLOAD

Diane L. Damos
Department of Psychology
Arizona State University, Tempe, AZ, 85287

Human factors practitioners often are concerned with mental workload in multiple-task situations. Investigations of these situations have demonstrated repeatedly that individuals differ in their subjective estimates of workload. These differences may be attributed in part to individual differences in definitions of workload (Hart, Childress, and Hauser, 1982). However, after allowing for differences in the definition of workload, there are still unexplained individual differences in workload ratings. The general purpose of the two studies reported in this paper was to examine the relation between individual differences in multiple-task performance, subjective estimates of workload, information processing abilities, and the Type A personality trait.

EXPERIMENT I

To date, two systems have been developed to classify individuals on the basis of their multiple-task performance. One of these was developed by R. S. Kennedy and dichotomizes individuals on the basis of their performance on an auditory monitoring task. One group, the broad bandpass group, shows little or no decrement with time on the complex version of the monitoring task but the typical vigilance decrement on the simpler version of the task. The second group, the narrow bandpass group, shows no vigilance decrement on the simple version of the monitoring task but performs poorly on the complex version.

The second system was developed by Damos (Damos and Wickens, 1980; Damos, Smist, and Bittner, 1983) and classifies individuals on the basis of the response strategy they use to perform a memory-classification task combination. Four strategies have been identified: simultaneous, alternating, massed, and mixed. Subjects using a simultaneous response strategy respond to stimuli from both tasks within some small interval (typically less than 30 ms). Those using the alternating strategy make one response at a time to each task, alternating between

PRECEDING PAGE BLANK NOT FILMED

the two. If more than two consecutive responses are emitted consistently to one task before responding to the other, the strategy is classified as massed. A few subjects may use all three strategies within a given trial or may use different strategies on different trials. These subjects are referred to as mixed.

The primary purpose of this study was to examine the relation between Kennedy's and Damos' classification systems and to determine if there were any consistent between-group differences in subjective estimations of workload or information processing abilities. Three basic tests of information processing ability were selected for use: digit span, hidden figures, and time estimation.

Method

Tasks. Because detailed descriptions of the tasks are given elsewhere (Damos, 1984), each task will be described only briefly. For the memory task the subject responded to the stimulus preceding the one currently displayed. The stimuli consisted of the numbers one through four and the subject pressed one of four keys with her right hand. For the classification task two numbers were displayed concurrently. The numbers could differ in size (big or little) and name (e.g. five versus six). The subject determined the number of dimensions on which the stimuli were alike (zero, one, or two) and pressed one of three keys with her left hand.

For the auditory monitoring task the subject listened to a series of high, medium, and low tones for 60 min. In the simple version the subject monitored only the low tone. After the fourth occurrence of the low tone, she pressed a button and began counting again. In the complex version she monitored all three tones simultaneously but independently.

A semi-automatic digit span test was used in this study. The subject saw a sequence of numbers. At the end of the sequence the subject repeated the numbers in order to the experimenter. The hidden figures test (Ekstrom, French, Harmon, and Derman, 1976) measured field independence. The time estimation test required the subject to estimate the passage of 10 s. When the subject thought 10 s had passed, she pushed a key and began estimating 10 s again immediately. Each subject estimated ten 10-s intervals.

Ten bi-polar adjective workload scales were used. These scales were: overall workload, task difficulty, time pressure, actual performance, comfort level, mental/sensory effort, achievement performance, skill required, fatigue, and stress level. The scales were presented singly on sheets of paper with the scale itself represented by a vertical line 110 mm long. Subjects made a rating by drawing a line horizontally across the vertical line. Ratings were computed as the distance from the bottom of the vertical line to the mark.

Subjects. Thirty right-handed female subjects completed the experiment. They were paid for their participation.

Procedure. This experiment required approximately 1.25 h per day on each of four consecutive days. The subject performed the information processing tests on Day 1 and became familiar with the workload scales. On Day 2 the subject performed one version of the tone monitoring task. Half of the subjects performed the simple version on Day 2; half, the complex version. Day 3 was devoted to performing the memory and classification tasks used to assess response strategy. Finally, on Day 4 the subject performed the other version of the tone monitoring task. On all four days the subject rated each task on the ten workload scales immediately after she had completed it.

Results

Only the most important results will be discussed (for a more detailed description of the data, see Damos, 1984). One of the major purposes of this study was to examine the relation between Kennedy's and Damos' classification systems. However, no evidence of the broad bandpass/narrow bandpass distinction was found in the data and no analyses were conducted using this classification system. Subjects' response patterns on the memory-classification combination were analyzed using a technique described in Damos et al (1983). Ten subjects subsequently were classified as simultaneous responders, eight as alternators, seven as massed, and five as mixed.

Digit span and time estimation were significantly correlated (r=.38, p<.05). Consequently, a multivariate analysis of variance was conducted on the data. There were no significant between-group differences (p>.05). However, there were some interesting trends in the data with 10 s mean estimation scores ranging from 10.4 s for the simultaneous group to 7.7 s for the massed group. Digit span ranged from 7.5 for the mixed group to 6.5 for the massed group.

All ten workload scales showed a significant effect of task ($p<.05$). Two scales, overall workload and task difficulty, also showed significant between-group differences (F (3,26,)=3.06, p=.05 and F (3,26)=3.29, p=.04, respectively) with the mixed group giving the highest ratings.

Discussion

The major findings of interest in this experiment are the between-group differences in digit span and time estimation although these differences are not statistically significant. These results indicate that between-response strategy group differences may be related to differences in short-term memory, an idea which is explored in more detail in Experiment II. The results also show some between-group differences in the subjective experience of workload. These differences also are examined further in the second experiment.

EXPERIMENT II

The primary purpose of this experiment was to explore the relation between response strategy groups, measures of information processing ability, and subjective estimates of workload. Additionally, a measure of the Type A personality trait was included to determine if there was any relation between Type A behavior, which is marked by self-imposed time pressure, and the subjective experience of workload.

Method

Tasks. To identify the subject's response strategy, the memory and classification tasks described in Experiment I were used in this investigation. The digit span task and the time estimation task from Experiment I also were used without modification in this study.

Two additional information processing tasks were included in this study. One was a choice reaction time task performed at 1, 2, and 3 bits of information. Stimuli for this task were visually presented digits; the subject responded manually. The second was a memory search task, which used visually presented letters. Each subject performed this task using positive set sizes of three, four, and five letters. The subject also performed the time estimation task concurrently with the choice reaction time at the 3-bit level and with the memory search task using a positive set size of five letters.

Additionally, the subjects performed a matrix task and a mental arithmetic task alone and together. For the matrix task 5 by 5 matrix grids with five randomly selected illuminated cells were presented sequentially to the subject. The subject's task was to determine if the current matrix was a rotated version of the immediately preceding matrix or a different matrix. The subject responded "same" or "different" by pressing one of two keys with her left hand. For the mental arithmetic task randomly selected digits between zero and eight were presented sequentially to the subject. The subject indicated the absolute difference between the most recently displayed digit and the preceding digit using her right hand.

All of the bi-polar workload scales from Experiment I except comfort level and skill required were used in this study. Additionally, all of the subjects completed a modified version of the Jenkins Activity Survey to measure the Type A personality trait.

Subjects. Thirty right-handed female subjects completed the investigation. All were paid volunteers and none participated in Experiment I.

Procedure. This study required approximately 1.5 h on each of three successive days. On Day 1 the subject performed the digit span task and completed the modified Jenkins Activity Survey. She also performed the memory search task. On Day 2 the subject performed the choice reaction time task followed by the matrix task and the mental arithmetic task, which were performed alone and together. Day 3 was devoted to the memory and classification tasks and to time estimation. The subject first performed two trials of the time estimation task alone. Subsequently, she performed the time estimation task concurrently with the memory search task for two trials and with the choice reaction time task for two trials. After completing each of the tasks mentioned above, the subject rated her subjective experience of workload using the scales described previously.

Results

Only the statistically significant results will be discussed. Subject's response strategies were identified using the procedure described in Damos et al (1983). Three subjects used the alternating strategy; ten, the simultaneous strategy; and 15, the massed strategy. The remaining two subjects used the mixed strategy in this investigation. These two subjects' data were combined with those of the massed strategy subjects for the statistical analyses.

Statistically significant between-response group differences were found on the intercept of the memory search task (F (2,27)=5.78, p=.01). The intercept for the simultaneous group was 308 ms; for the alternating group, 490 ms; and for the massed group, 439 ms. Although the groups did not differ significantly on their single-task time estimation scores (p>.05), they did differ significantly on both dual-task combinations. The estimates for the three groups were 10.4 s for the simultaneous group, 7.8 s for the alternating group, and 13.2 s for the massed group when the time estimation task was done concurrently with the choice reaction time task (F (2,27)=3.42, p=.05). When time estimation was done concurrently with the memory search task, the estimates were 10.2 s for the simultaneous group, 10.7 s for the alternating group, and 14.1 s for the massed group (F (2,27)=5.48, p=.01).

All of the subjective workload scales except the frustration scale showed a significant effect of task (p<.05). However, none showed a significant effect of group (p>.05).

The seven subjects having the highest scores on the modified Jenkins Activity Survey (the top 23% of the sample) were categorized as Type A's and the nine subjects with the lowest scores (the bottom 30% of the sample), as Type B's. A two-way (group by task) ANOVA conducted on the subjective workload scales revealed a significant group by task interaction on the frustration scale (F (7,98)=2.81, p=.01). This interaction reflects differences in perceived frustration under dual-task conditions; Type A's rated their frustration much higher under dual-task conditions than Type B's. However, under single-task conditions there was no apparent difference between the Type A's and the Type B's.

Several other statistically significant differences between Type A's and Type B's were found. A two-way ANOVA (group by trial) performed on the time estimation scores of the choice reaction time-time estimation combination found a significant group by trial interaction (F(1,14)=8.27, p=.01). Type B subjects lengthened their estimation of 10 s from 10.9 s on the first trial to 13.1 s on the second. In contrast Type A subjects decreased their estimation from 12.5 s to 12.0 s over the two trials. On the choice reaction time task both groups of subjects improved about 200 ms over the two trials. However, the average reaction time for the Type A's was 806 ms; that for the Type B's, 954 ms.

Only the two-way ANOVA performed on the memory search data of the memory search-time estimation combination showed a significant between-group difference (F (1,14)=4.57, p=.05). Again, both groups improved with practice, but the average performance of Type A's was better than Type B's (756 ms versus 1004 ms).

Discussion

The trends in between-strategy group differences in digit span and time estimation found in Experiment I were not replicated in this study. However, there were statistically significant differences in dual-task time estimation scores on both of the time estimation combinations with the massed subjects having consistently the worst performance. Thus, it appears that between-group differences in time estimation occur only under dual-task conditions. The underlying cause of these differences is not apparent and awaits further research.

Probably the most interesting results obtained in these two investigations concerns the differences between Type A and Type B subjects. Type A individuals have been characterized as hard-driving competitive people with a sense of time urgency. Their competitiveness appears to be reflected in their superior dual-task performance on both time estimation combinations but is also evident in their higher frustration with these tasks. Again, the underlying source of these differences is not apparent and should be explored more thoroughly.

REFERENCES

Damos, D. Individual differences in multiple-task performance and subjective estimates of workload. Paper under review, Perceptual and Motor Skills.

Damos, D., Smist, T., and Bittner, A., Jr. Individual differences in multiple-task performance as a function of response strategy. Human Factors, 1983,25, 215-226.

Damos, D. and Wickens, C. The identification and transfer of timesharing skills. Acta Psychologica, 1980, 46,15-39.

Ekstrom, B., French, J., Harman, H., and Dermen,

D. Manual for Kit of Factor-Referenced Cognitive Tests. Princeton, New Jersey: Educational Testing Service, Office of Naval Research Contract N0014-71-C-0117, 1976.

Hart, S., Childress, M., and Hauser, J. Individual definitions of the term "workload." Paper presented at the Psychology in the DOD Symposium, Colorado Springs, Colorado, 1982.

Mental Models

N85
14544

UNCLAS

D9

N85 14544

MENTAL MODELS OF INVISIBLE LOGICAL NETWORKS

Penelope Sanderson

Departments of Psychology and Industrial Engineering
University of Toronto
Toronto Canada M5S 1A5

ABSTRACT

This experiment required subjects to discover the structure of a logical network whose links were invisible. Network structure had to be inferred from the behaviour of the components after a failure. It was hypothesised that since such failure diagnosis tasks often drew on spatial processes, a good deal of spatial complexity in the network should affect network discovery. Results showed that the ability to discover the linkages in the network was directly related to the spatial complexity of the pathway described by the linkages. This effect was generally independent of the amount of evidence available to subjects about the existence of the link. These results raise the question of whether inferences about spatially complex pathways were simply not made, or whether they were made but not retained because of a high load on memory resources.

INTRODUCTION

There is currently much interest in the idea of a 'mental model' - a hypothetical construct which describes the form of private knowledge about a task or a system. It is generally agreed that the mental model is usually an approximation to the reality of a task or system (Gentner and Stevens, 1983). A good deal of research has been devoted to finding out when the fact that the mental model is an approximation will be evident, and when not. Such work suggests that mental models draw upon visual, verbal and kinaesthetic codes and that multiple mental models can co-exist (Eberts, 1982; Wickens and Kessel, 1981; Rasmussen, 1979).

Mental models can also be seen as private 'theories' as to what mediates between the input and output of systems whose structure is not necessarily explicit. This characterisation is important for an understanding of how the contemporary human process control operator controls and troubleshoots a complex, slowmoving system. From time to time there have been calls for the development of reliable, repeatable, generalisable and quantifiable laboratory tasks to study such performance at a theoretical level (John, 1957; Coates, Alluisi and Morgan, 1971; Rouse, 1978). The result has been a family of tasks which borrow heavily from mechanics and electronic engineering (Pylyshyn, 1963; Alluisi and Coates, 1967; John, 1957; Pikaar and Twente, 1981). The majority of these studies have focused on the subject's ability to discover the principles behind a task and the rules governing its input-output relationships. These principles and rules have to be inferred by the subject from the response of the system to a series of inputs. The subject's moment-by-moment understanding of the principles and rules can be said to be his current 'mental model' of the task.

The research to be reported here is in this tradition. However in its methodology it borrows from fault diagnosis studies which have made the principles behind the task more explicit and which have focused on aids to the achievement of performance goals. In this family of tasks are the recent thorough experiments by Rouse (1981) using the Task and Fault procedures, and similar experiments by Brooke, Cook and Duncan (1983) and Goldbeck, Bernstein, Hillix and Marx (1957). In Rouse's Task procedure, subjects are confronted with a linked network of logical components. A failure somewhere in the network leads to at least some of the components in the right-hand column taking on values of 0 rather than 1. Subjects have to discover which component has failed, causing these 0 values. They do this by asking the computer for the status of various links until they can correctly hypothesise which component caused the failure.

The present research will focus on how subjects might learn the structure of a network of this kind. An experimental problem similar to Rouse's Tasks One and Two will be used, but with two important differences. First, only one network will be used in the experiment. Its underlying linkages will never change at any point in the procedure. Second, the linkages in this network will always be invisible, so that the subject has no information at all about network structure at the beginning of the experiment. Thus, the only clues to network structure will be the effect that individual failures have on the total network. Evidence about structure will have to be collected over many trials and many failures, to be integrated and held in memory by the subject. The result of this process may be thought of as the subject's 'mental model' of the network.

If the subject relies entirely upon logic to discover the structure of this invisible network, then there are three general types of logical operation available which will achieve this (these will be outlined below). However it is well-known in both the psychological and human factors literature that subjects often fail to exploit logic to its full extent (Kahneman, Slovic and Tversky, 1982; Mynatt, Doherty and Tweney, 1977; Wason, 1960; Bainbridge, 1981; Rasmussen, 1981; Rouse, 1978; Hunt and Rouse, 1980). This usually happens when the perceived complexity of the problem overwhelms the subject's ability to use logic, or when the representational form of the problem triggers a preference for an alternative heuristic. In a human factors environment it is important to be able to predict when this will happen.

There are two lines of evidence which combine to suggest that in the proposed network discovery task, the representational form of the task may interfere with subjects' logical inferences, particularly as subjects will be under a high memory load. The first line of evidence is that failure detection and diagnosis tasks have been found to be highly dependent on spatial factors. For example, Landeweerd (1979) reports that the ability to draw a visual schematic of a distillation process correlates with failure diagnosis performance but not with control performance. Moreover, Weingartner (1982) has attacked this issue with dual task methodology. With multi-element failure detection as the primary task, a spatial secondary task provided more interference than a verbal secondary task. Therefore it seems that spatial factors are at work in fault diagnosis tasks. The second line of evidence concerns the variables which contribute to the perceived complexity of spatial patterns. Research shows that judgements of pattern complexity are related to such factors as the variability of angular changes and the number of turns in the contour of a shape (Attneave, 1957; Lemay, in Rouse and Rouse, 1979), the degree to which the apparent consistency of a pattern is violated (Hake and Eriksen, 1956) and the number of alternatives a pattern has (Garner, 1970). Given these results, it should be possible to position network components in such a way that some links form perceptually more complex pathways than others. If the network discovery task taps into spatial processes and if some linkages trace out spatially simple paths and others spatially complex paths, then subjects' use of logic may well be influenced by the spatial complexity of the linkages they are trying to discover.

Accordingly, in the present experiment the subjects will always perform the network discovery task with the same underlying network. This is seen in Figure 1. However this network will be displayed to subjects in two spatial configurations, Pattern A and Pattern B (the linkages shown in Figures 2 and 3 are of course not displayed to subjects). It can be seen in the underlying network that there are two pathways from component 2 (#2) to #8 - one through #6 and the other through #4. In pattern A the #2-#6-#8 pathway is straight while the #2-#4-#8 pathway is more circuitous. Pattern B has been designed so that the opposite is true. Moreover, in pattern A the #3-#6-#8 pathway is straight while the #1-#4-#8 pathway reverses direction. Again, in pattern B the opposite is true. In all these cases, it is hypothesised that the more circuitous route will be the harder one for subjects to discover. Discovery will be assessed by subjects' ability to include the link in a drawing of the network.

An examination of the underlying network in Figure 1 shows that if any of components #1, #2, #3, #4, #6, #8, #9 or #10 fails, component #10 will carry a signal of 0. In two conditions, each of these components had an equal probability ($p=.125$) of failure (conditions A-Equal and B-Equal). In the other two conditions, the probability of failure for each component was chosen so as to bias subjects towards discovering the links that the patterns might make difficult. For pattern A it was hypothesised that the #2-#6 and #6-#8 links might be hard to find. So, in the biased condition #2, #3 and #6 were made to fail more frequently than #1 and #4, emphasising the harder pathway ($p=.1875$ vs $p=.0625$, condition A-Biased). A similar manipulation was made for pattern B. Here, the #2-#4 and #4-#8 links might be hard to find, so #2, #1 and #4 were made to fail more frequently than #3 and #6 (B-Biased). If the subjects are using logical processes, then sheer weight of evidence may overcome the difficulty with the harder links.

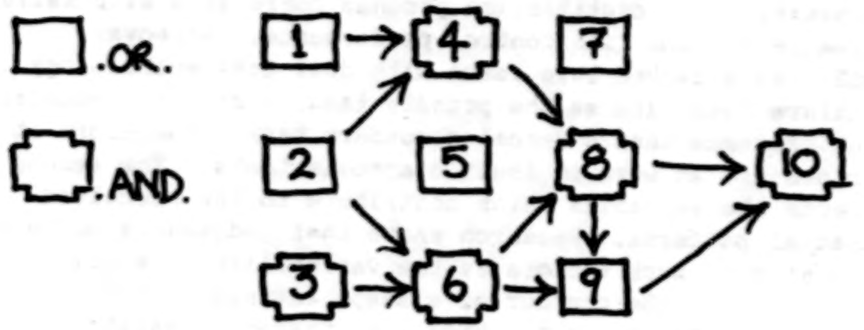

.OR.
.AND.
1
4
7
2
5
8
10
3
6
9

FIG 1: The network in its neutral configuration

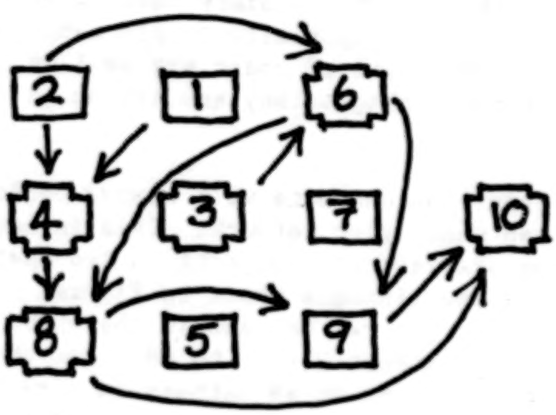

2
1
6
4
3
7
10
8
5
9

FIG 2: Structure of Pattern A

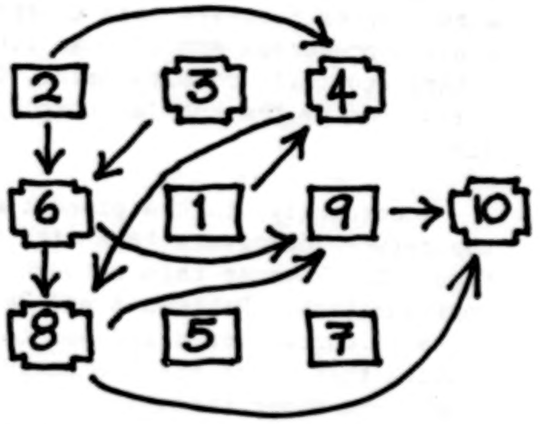
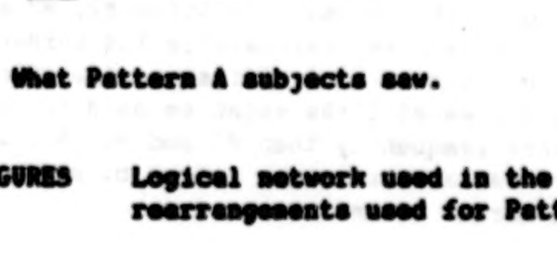
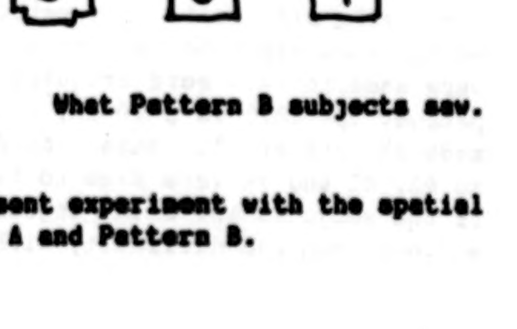

2
3
4
6
1
9
10
8
5
7

FIG 3: Structure of Pattern B


1
4
7
2
5
8
10
3
6
9

What Pattern A subjects saw.


1
4
7
2
5
8
10
3
6
9

What Pattern B subjects saw.

FIGURES Logical network used in the present experiment with the spatial rearrangements used for Pattern A and Pattern B.

SIMULATION AND IDEAL SEARCH STRUCTURE

There are three general types of logical operation available which will recover the structure of this network. First, if two components, X and Y, are seen with a status of 0 and the subject learns that X is the failure, then Y must be later in the sequence from X and must have a direct or indirect link from it. It can also be concluded that Y does not lead to X. Second, it is possible to use inductive logic to eliminate connections to .OR. components. Third, redundant links can be 'housecleaned'. If X leads to .AND. component Y and Y leads to .AND. component Z, then an X to Z link is redundant and can be discarded.

To test whether these three rules were sufficient to discover the structure of the network, a simulation was run, named PLAULTER. PLAULTER 'played' the fault diagnosis task the same way as a human subject, choosing views and making hypotheses. At the end of each trial when the failure was known, PLAULTER collected information about the structure of the network using the three operations described above, and added a code for this information to an internal reachability matrix. PLAULTER was put to work on a wide variety of networks which differed in structure and in size from 4 to 25 components. As long as there was no feedback in the network, PLAULTER was always able to discover its structure. PLAULTER was also used in the data analysis, monitoring subjects' performance in order to make inferences about the maximum amount of logical information they had about the structure of the network at various points over their session.

The best long-term expected number of views that a subject can achieve for the network used in this task is 3.125 views, the maximum for any failure being 4 views and the minimum 2 views. This is achieved if the subject consistently follows a 'half-split' strategy, eliminating as close as possible to half the remaining potential failure sources on each view. There are many search strategies which, if followed consistently, will lead to this result, but many other less satisfactory strategies which lead to a greater expected number of views.

METHOD

Design

The factors to be manipulated in this experiment were, first, the patterning of network isomorphs ('Pattern' A and B) and, second, the probability with which individual components failed ('Probability of Failure' Equal or Biased). A between subjects design was used, with 8 subjects in each condition.

Failures occurred in exactly the same sequence for the A-Equal and B-Equal conditions, so that potential evidence about the network was available at the same time, regardless of condition. In the conditions A-Biased and B-Biased, at given points a low probability failure in this sequence was replaced by a high probability failure, yielding slightly different sequences. Still, as in the Equal conditions, all subjects within a Biased condition experienced the same sequence of failures.

Procedure and Instructions

Subjects see a collection of numbered logical components on the video screen. They are told that the components are linked in a fixed pattern to make a network, but that these links will never be visible. Instead, they will have to form some kind of a mental representation of how the components are linked. Some of the components are logical .AND. gates and some are logical .OR. gates and this distinction is indicated by the shape of the component. The on(1)/off(0) status of components 1 to 9 is not visible to the subject at this stage. However the status of component 10 can always be seen as it signals whether the final outcome of the network is 1 or 0.

Subjects are told that when the network is behaving normally, all components carry a signal of 1 and that component 10 has a value of 1. However, if there is a failure somewhere in the network, one of the components will drop from a value of 1 to a value of 0. The network is structured so that if there is a failure, it will eventually filter through to component 10, which will be seen to take on a value of 0. Once this happens, the subject has to discover the component responsible for the failure. To do this, it is necessary to see which components are sending signals of 1 and which are sending signals of 0 on this trial, as all components sending signals of 0 are potentially responsible for the network failure.

The subject asks to see the status of a component by typing the letter 'V' (View) followed by the number of the component in question. The border of the component will then change colour. If the component is sending a signal of 1 the border becomes thick and bright and if it is sending a signal of 0 the border becomes thick and dark. The subject can view the status of as many components as he or she wishes. At any stage the subject may wish to make a hypothesis that a certain component is responsible for the network failure. To do this the subject types 'H' (Hypothesis) followed by the component number. If the hypothesis is wrong the word 'WRONG' appears and the subject can to do more viewing and hypothesising. If the hypothesis is correct the word 'CORRECT' appears and no more viewing and hypothesising can be done. At this point the subject can examine the display on the screen for as long as required, before going on to the next trial. When the command to continue is given, all the components are restored to their neutral colours, the failure is 'repaired' and component 10 can be seen to carry a signal of 1 again. However then a failure occurs somewhere else in the network and subjects repeat the fault diagnosis procedure.

Subjects were told that components could have either 0 or 2 inputs and 0, 1 or 2 outputs, so that some components would have inputs but no outputs ('dead-ends'), some outputs but no inputs ('originators'), some both inputs and outputs ('normal') and some neither ('dodoes'). The .AND. and .OR. logic was explained to them with a diagram and the diagram was available to them throughout the experiment. They were given three goals to work towards: (1) learn how the network is structured, (2) find the failed component in as few views as possible and (3) the first hypothesis made must be the correct hypothesis.

The experimental session was two hours long. Instructions took about 15 minutes and then the subject was allowed to proceed with the fault diagnosis task. Subjects were not allowed to take notes or make drawings while they did the task. However, at 35-minute intervals (one-third and two-thirds through the session and at the end) the subject was interrupted and given a picture of the network as it appeared on the screen. They were asked to draw in where they currently thought the links were, based on the evidence they had seen rather than on guesses about where they might be. When the subject was satisfied with the drawing, generally after about 3 minutes, they described it to the experimenter. Then the drawing was taken away and the subject was allowed to proceed with the fault diagnosis task.

Subjects

Subjects were 32 undergraduate students from University of Toronto and were paid $10 each for participating. They were randomly assigned to the four conditions until there were 8 subjects per condition. Students who had taken intensive courses in computer science or electrical engineering were not used in the experiment. No subject had experienced the network discovery task before.

RESULTS

FAULT DIAGNOSIS

If a subject completed fewer than 50 trials, or was still hypothesising at random by the end of the experiment, his or her performance was not comparable with that of other subjects and could not be included. The data for six out of 38 subjects were discarded using this rule. For the 32 subjects whose data were included, the overall average number of trials completed was 106.25 (SD=33.1). A two-way ANOVA with factors Pattern (A/B) and Probability of Failure (Equal/Biased) gave no evidence to suggest that the number of trials completed was different under different experimental conditions.

Summary statistics

All subjects in all conditions showed considerable progress towards the goals set in the instructions. Overall performance for each subject on the first 50% of trials ('first half') was compared with performance on the second 50% ('second half') for views and hypotheses. Two three-way ANOVAs were used, the factors being Pattern, Probability of Failure and Half. In both cases, the main effect of Half was significant. All subjects took fewer views in the second half before finding the failure (F(1,28)= 194.9, p<.001: views in 1st half=5.17, 2nd half=3.54). On nearly all occasions, in the second half their first hypothesis was the correct hypothesis (F(1,28)= 15.11, p<.001: hypotheses in 1st half=1.36, 2nd half=1.12). For both views and hypotheses, no other main effects or interactions were significant.

Transition matrices were made from the raw data of the subject's moves. Given that a subject had just viewed or hypothesised a certain component, with a given result, probabilities of the next action could be calculated. From the transition probabilities, it was possible to discern the development of search strategies. For 30 of the 32 subjects, a single clear, idiosyncratic strategy had emerged by the end of the experiment. One result of this was that the entropy in the subjects' transition matrices for the first 32 trials was much higher than for the last 32 trials, by which time search was highly structured. The entropy for the first 32 trials was .71 and for the last 32 it was .48, compared with a maximum possible of approximately .8 and a minimum of approximately .4. A three-way ANOVA found that this difference was highly significant ($F(1,28)=534.6$, $p<.001$), and no main effects or interactions with the experimental conditions emerged. The general impression from the results is that subjects start by viewing components exhaustively and not in any particular order. The final search strategy emerges rather suddenly and generally the subject will not alter it for the rest of the experiment.

Hypothesised Effects of Pattern and Probability of Failure

For 32 subjects, 25 different search patterns were generated, making it impossible to find a consistent preference for one search strategy over another between conditions. However, condition did appear to have some effect on the initial component viewed. Components #9 and #8 can be classified together as the components which link directly to #10, components #6 and #4 as having both inputs and outputs within the network proper and components #1, #2 and #3 as being 'originators' or sources. The results show that #9 and #8 were more frequently the starting components in the two Equal conditions (Equal: 12 out of 16, Biased: 6/16) while #4 and #6 were more frequently the starting components in the Biased conditions (Equal: 4/16, Biased: 8/16), but this trend failed to reach significance. However, more subjects achieved the optimal search (average of 3.125 moves) in the two Biased conditions (8/16) than in the two Equal conditions (2/16) ($\chi^2(1)=4.69$, $p<.05$).

NETWORK DRAWINGS

Subjects' first, second and third network drawings were scored for the presence or absence of real and non-existent links. If a subject omitted a real link or included a non-existent link on one or more drawings, he or she was said to have 'difficulty' placing or eliminating that link. For all subjects the third drawing was considerably more accurate than the first drawing. While the final drawing was often very close to the actual network, only 2/8 subjects in each of the four conditions produced a perfect final drawing.

Before looking at the influence of conditions on the ability to draw the network, it is important to know whether the overall difficulty with the drawings was the same for each condition. It seems that this was the case. Three different measures of the frequency of difficulty were examined in 2x2 contingency tables with the Pearson chi-squared test of independence. If a subject showed difficulty with a link, this was added to the frequency count and the total was the cell entry. The results are shown in Table 1. First, the total number of link omissions was not significantly different over conditions. Second, the total number of times omissions were still found by the final drawing was not significant. Finally, the total number of 'commissions' of false links was not significantly different over conditions. Thus any systematic differences over conditions in the ability to draw links cannot be attributed to the general difficulty of the condition, but instead are specific interactions of link difficulty with condition.

The spatial Pattern manipulation appears to have exerted quite an influence on which links subjects found difficult. This was usually true even when evidence for the links was made more available in the Biased conditions, thus favouring the null hypothesis over the experimental hypothesis.

As hypothesised, the Pattern A subjects had considerable difficulty with the #2-#4 links while the Pattern B subjects tended to have more difficulty with the #2-#6 links (see Table 2). In a chi-squared test of independence there was a significant dependency between Pattern and the difficulty of the two links ($\chi^2(1)=4.71$, $p<.05$). In the last drawing made, many subjects had still not found the link supposed to be harder despite the inclusion of the Biased conditions and this difficulty still showed borderline significance ($\chi^2(1)=3.02$, $p<.1$). If we remove the Biased conditions and look at difficulty in the last drawing for the Equal conditions only, the overall difficulty shows borderline significance ($\chi^2(1)=2.86$, $p<.1$) and this finding becomes more strongly significant for links never found ($\chi^2(1)=3.99$, $p<.05$). In general, the Bias manipulation appears to have had very little influence in reducing the difficulty subjects had in finding the spatially more complex link.

It was also hypothesised that subjects working on Pattern A would find the #6-#8 link harder to find than the #4-#8 and that the reverse should be true for subjects working on Pattern B. As Table 3 shows, this hypothesis was not upheld when Equal and Biased conditions were combined and showed only a trend in the hypothesised direction when the Biased conditions were eliminated. However a related effect emerged in the results. This effect concerned the fact that in Pattern A the #3-#6-#8 pathway reversed direction and the #1-#4-#8 was straight, while the opposite was true for Pattern B. There was a significant tendency for subjects using Pattern A to hypothesise the non-existent link #3-#8 more frequently than #1-#8, and for subjects using Pattern B to hypothesise #1-#8 more frequently than #3-#8 (see Table 4). Collapsing over the Biased and Equal conditions this was significant ($\chi^2(1)=6.52$, $p<.025$) and it was also significant for the Equal condition alone ($\chi^2(1)=6.48$, $p<.025$). In each case, the preferred but non-existent link was an abbreviation of the more complex pathway that really existed. The effect was not significant in the Biased conditions.

TABLE 1. NUMBER OF LINKS FOUND DIFFICULT, SUMMED OVER SUBJECTS

	Difficulties at any stage		Links never found		False links drawn	
	ALL A	ALL B	ALL A	ALL B	ALL A	ALL B
EQUAL	33	42	13	14	16	13
BIASED	45	41	14	12	21	13
	n.s.		n.s.		n.s.	

TABLE 2. SUBJECTS SHOWING DIFFICULTY WITH #2-#4 AND #2-#6

	Difficulty at any stage		Link never found		Difficulty at any stage		Link never found	
	ALL A	ALL B	ALL A	ALL B	A-EQU	B-EQU	A-EQU	B-EQU
#2-#4	4	12	2	7	2	8	0	4
#2-#6	15	9	6	3	8	4	4	2
	$\chi^2(1)=4.71*$		$\chi^2(1)=3.02$		$\chi^2(1)=2.86$		$\chi^2(1)=3.99*$	

TABLE 3. SUBJECTS SHOWING DIFFICULTY #4-#8 AND #6-#8

	Difficulty at any stage		Link never found	
	ALL A	ALL B	ALL A	ALL B
#4-#8	7	9	1	0
#6-#8	10	8	5	2
	n.s.		n.s.	

TABLE 4. SUBJECTS DRAWING FALSE LINKS BETWEEN #1-#8 AND #3-#8

	ALL A	ALL B	A-EQU	B-EQU	A-BIAS	B-BIAS
#1-#8	1	8	0	4	1	4
#3-#8	8	3	6	1	2	2
	$\chi^2(1)=6.52*$		$\chi^2(1)=6.48*$		n.s.	

The only strong effect of Bias on subjects' ability to draw network links was seen for the links for which there was infrequent evidence. Link #1-#4 was less frequent and more difficult in the A-Biased condition and #3-#6 was less frequent and more difficult in the B-Biased condition ($\chi^2(1)=4.99$, $p<.05$).

The link which eluded subjects the most, regardless of condition, was #8-#9. Twenty-two out of 32 subjects experienced difficulty with it and 14/32 had still not found it by the end of the experiment. This may to be because #9 was the only .OR. component with inputs, or because #9 is in a rather difficult spatial and logical configuration.

Relationship of Network Drawing with Fault Diagnosis Performance

It is important to establish the relationship between the evidence subjects saw and the inferences about structure that they were able to make. On the one hand, the spatial characteristics of the task may have drawn subjects to provide themselves with biased information. On the other hand, the spatial characteristics may have led them to ignore the information that was present.

It is possible that Pattern A subjects saw the evidence for link #2-#6 on a much later trial than #2-#4, and that the reverse was true for Pattern B subjects. If so, the omission of the harder link in at least some of the drawings would be due to lack of evidence. However the data suggest otherwise. As Table 5 shows, nearly all subjects saw the two pieces of evidence on the same trial. Thus it seems that not enough subjects saw one link before the other to make the difficulty with the drawings entirely attributable to lack of evidence.

It may also be the case that subjects provided themselves with the evidence for the spatially easier link more frequently than evidence for the harder link and this made them overlook the harder link. However, if anything the opposite is true, as Table 6 illustrates. Thus the evidence for inferring the spatially harder link was definitely available.

A similar approach can be taken for the non-existent #1-#8 and #3-#8 links. In all conditions except B-Biased, evidence for the fact that #1-#8 was not a link but rather the ends of the #1-#4-#8 pathway always came before the evidence that #3-#8 was the ends of the #3-#6-#8 pathway. Difficulty with the link is quite independent of this factor. Difficulty is also independent of the number of times subjects saw the complete pathway after the evidence against the link had become available.

Landeweerd's (1979) results suggested that ability at fault diagnosis and at drawing an image of the system were correlated, as both tasks drew on spatial processes. That finding was partially echoed here. The total number of mistakes that subjects made on the drawings was examined for correlations with fault diagnosis measures. First, significant correlations were found between mistakes and entropy in the transition matrices for the first 32 trials ($r(30)=.67$, $p<.001$) and the last 32 trials ($r(30)=.61$, $p<.001$).

TABLE 5. SUBJECTS SEEING EVIDENCE FOR #2-#4/#2-#6 IN SAME TRIAL

	A	B
EQUAL	7	8
BIASED	6	6

TABLE 6. SUBJECTS SEEING EVIDENCE FOR EASIER/HARDER LINK FIRST

	A-EQUAL	A-BIAS	B-EQUAL	B-BIAS
EASIER FIRST	2	3	4	1
HARDER FIRST	6	5	4	7

TABLE 7. NUMBER OF SUBJECTS ACHIEVING OPTIMAL SEARCH STRATEGY

	ALL A	ALL B	
3.125 (OPT)	2	8	
3.25-4.125	14	8	$\chi^2(1)=4.69$*

Second, no correlation was found between the total number of mistakes on the drawing and each subject's expected average number of views ($r(30)=.22$). Point biserial correlations were calculated between whether the subject found the harder link from #2 and fault diagnosis performance measures. There were minor but significant correlations with the actual (not expected) average number of views for the last 32 trials ($r(30)=-.39$, $p<.05$) and with transition matrix entropy for the last 32 trials ($r=-.4$, $p<.05$).

The data made it clear that knowledge about the harder link was not a necessary condition for achieving the most efficient search strategy. Of the 10/32 subjects whose search strategy had an expected average number of moves of 3.125, 2 subjects had still not found the most difficult link by the end of the experiment. Nor did knowledge about the harder link guarantee an efficient search strategy. Of the 19/32 subjects who eventually found the harder link, only 6 had a search strategy whose expected average was 3.125.

DISCUSSION

The results of this experiment suggest that a subject's ability to make logical inferences about network links is affected by the spatial complexity of the link in question. Plausible linkages or simple spatial configurations were apparently easier to find or more likely to be invented. In the Equal conditions, the #2-#4 link was easier to identify than #2-#6 in Pattern A and the #2-#6 link was easier to identify than #2-#4 in Pattern B. This was so even though PLAULTER showed that the evidence for both links was usually available to subjects within the same trial. The existence of both links should have been deduced using the first type of logical deduction described (see above). Indeed, with the present results it is impossible to state categorically that an appropriate deduction for the harder link was not made at the same time as the easier link. It may have been noted, but not retained or reinforced sufficiently well to appear on the network drawing.

The fact that the easy links form a simple straight line with the #4-#8 link for Pattern A, and with the #6-#8 link for Pattern B, apparently overwhelms the possibility of other pathways. Moreover, the tendency for Pattern A subjects to draw a #3-#8 link and for Pattern B subjects to draw a #1-#8 link seems to reflect a similar preference for a spatially simple pathway. The preference for simple pathways is probably related to the workload of the task. The demands on working memory and intermediate memory stores are quite high, particularly at the start of the experiment. Under such conditions the developing mental model of the network clearly suffered.

The Probability of Failure condition has very little effect on apparent knowledge of the outputs from #2, especially for Pattern A where the effect of the configuration was most marked. However, an increased probability of failure of #3 and #6 in Pattern A and of #1 and #4 in Pattern B leads to more omissions of the less frequent links and fewer false hypotheses about the leads into #8. This demonstrates that in some cases at least, frequency of viewing evidence leads to better knowledge. It also poses the question of why the greater frequency of evidence should not then work for the harder outputs of #2.

It is always possible that the spatial bias may have been exaggerated by the fact that dual pathways exist between #2 and #8. Subjects may not have realised that seperate signals from a component could meet up again. However there are arguments against this interpretation. First, subjects' inability to find the spatially harder link generally meant that, in their drawings, there was an input missing to #6 in Pattern A and to #4 in Pattern B. Subjects knew this was impossible, as all components had to have 0 or 2 inputs. Many subjects expressed confusion about the source of the second input to this component, but the difficulty with the drawings persevered, despite this awareness. Second, the tendency for Pattern A subjects to draw #3-#8 and for Pattern B subjects to draw #1-#8 often generated a situation where seperate signals from #3 (Pattern A) or #1 (Pattern B) met up again at #8, via #6 and #4 respectively. Although this link was redundant in terms of the actual network, it shows that subjects often did allow this sort of structure in their drawings. However it is necessary to run a strong test of the spatial bias hypothesis where, for instance, the difficulty of the pathways of interest is controlled for by other dual pathways within the network.

The network drawing results may also be due to biases present at the time of drawing, rather than biases in what the subject knew about the structure of the network at that point. Convergent evidence is needed for the existence of a gap in the subject's knowledge. To a small degree, this is provided by the fact that the final failure search strategy for 11/16 Pattern A subjects did not include the transition "If #6 is 0, then view #2". Similarly, the final failure search strategy for 10/16 Pattern B subjects did not include the rule "If #4 is 0, then view #2". These figures are not compelling, but then they do reflect the final rather than an intermediate strategy. By this point almost half the subjects had found the link and may have modified their strategies. The bias in subjects' knowledge is also clear at the end of the experiment when they are shown the correct network structure to compare with their own drawings. If they had not found a particular link, subjects generally were surprised to see it on the diagram, claiming that they had never suspected its existence or had seen evidence against its existence, rather than that they had forgotten to draw it in. Future experiments should provide other convergent measures of knowledge in addition to the drawings.

In the light of Landeweerd's (1979) and Weingartner's (1982) work, it is interesting that correlations emerge between failure diagnosis performance and the ability to draw diagrams. It is particularly interesting that the correlations emerge not with the efficiency of the final failure search strategy but with the entropy in the transition matrix. This means that the more consistently the subject stuck to a search strategy, the less likely he or she was to omit links or hypothesise false links on the drawings. The actual efficiency of the search strategy was not crucial.

Future experiments are under development which should clarify whether the difficulty with links is due to reluctance in making logical inferences about some patterns over others, or to mental load associated with retaining such inferences. Subjects will be required to make on-line hypotheses about the presence or absence of links in more complex networks and verbal protocols will be taken. Assuming these manipulations do not radically alter the nature of the task, subjects' developing mental models of the network will be articulated more closely with their failure detection performance and with objective calculations of the evidence currently available.

REFERENCES

Alluisi, E.A. & Coates, G.D. (1967) A code transformation task that provides performance measures of nonverbal mediation (COTRAN). NASA Contractor Report No. CR-896.

Attneave, F. (1957) Physical determinants of the judged complexity of shapes. Journal of Experimental Psychology, 53, 221-227.

Bainbridge, L. (1981) Mathematical Equations or Processing Routines? In J. Rasmussen and W.B. Rouse (Eds.), Human Detection and Diagnosis of System Failures. New York: Plenum Press.

Brooke, J.B., Cook, J.F. & Duncan, K.D. (1983) Effects of computer aiding and pre-training on fault location. Ergonomics, 26, 669-686.

Coates, G.D., Alluisi, E.A. & Morgan, B.B. (1971) Trends in problem-solving research: twelve recently described tasks. Perceptual and Motor Skills, 33, 495-505.

Eberts, R. (1983) The development of an accurate internal model for high order systems. Proceedings, 26th Annual Meeting of the Human Factors Society. Seattle, WA: Human Factors.

Garner, W.R. Good patterns have few alternatives. American Scientist, 58, 34-42.

Gentner, D. & Stevens, A.S. (1983) Mental Models. Hillsdale NJ: Erlbaum Associates.

Goldbeck, R.A., Bernstein, B.B., Hillix, W.A. & Marx, M.H. (1957) Application of the half-split technique to problem-solving tasks. Journal of Experimental Psychology, 53, 330-338.

Hake, H.W. & Eriksen, C.W. (1956) Role of response variables in recognition and identification of complex visual forms. Journal of Experimental Psychology, 52, 235-243.

Hunt, R.M. & Rouse, W.B. (1980) Problem solving skills of maintenance trainees in diagnosing faults in simulated power-plants. Technical Report, University of Illinois at Urbana-Champaign

John, E.R. (1957) Contributions to the study of the problem-solving processes. Psychological Monographs, 71, 1-39.

Kahneman, D., Slovic, P. & Tversky, A. (1982) Judgement under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press.

Landeweerd, J.A. (1979) Internal representation of a process fault diagnosis and fault correction. Ergonomics, 22, 1343-1351.

Mynatt, C.R., Doherty, M.E. & Tweeny, R.D. (1977) Confirmation bias in a simulated research environment: An experimental study of scientific inference. Quarterly Journal of Experimental Psychology, 29, 85-95.

Pikaar, R.N. & Twente, T.H. (1981) How people discover input/output relationships. In H. Stassen (Ed.) First European Annual Conference on Human Decision Making and Manual Control. New York: Plenum Press.

Pylyshyn, Z. W. (1963) Search strategy and problem structure in heuristic problem solving. British Journal of Psychology, 56, 197-215.

Rasmussen, J. (1979) On the structure of knowledge - a morphology of mental models in a man-machine system context. Riso-M-2192 Report.

Rasmussen, J. (1981) Models of mental strategies in process control. In J. Rasmussen & W.B. Rouse (Eds.), Human Detection and Diagnosis of System Failures. New York: Plenum Press.

Rouse, W.B. (1978) Human problem solving performance in a fault diagnosis task. IEEE Transactions on Systems, Man and Cybernetics, SMC-8, 258-270.

Rouse, W.B. & Rouse, S.H. (1979) Measures of complexity of fault diagnosis tasks. IEEE Transactions on Systems, Man and Cybernetics, SMC-8, 720-727.

Rouse, W.B. (1981) Experimental studies and mathematical models of human problem solving performance in fault diagnosis tasks. In J. Rasmussen & W.B. Rouse (Eds.), Human Detection and Diagnosis of System Failures. New York: Plenum Press.

Wason, P.C. (1960) On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12, 129-140.

Weingartner, A. (1982) The internal model of dynamics systems: An investigation of its mode of representation. Undergraduate honours thesis, University of Illinois, Department of Psychology, Champaign, IL.

Wickens, C.D. & Kessel, C. (1981) Failure detection in dynamic systems. In J. Rasmussen & W.B. Rouse (Eds.), Human Detection and Diagnosis of System Failures. New York: Plenum Press.

N85
14545

UNCLAS

N85 14545 D10

The Representation Of Action Plans In Long Term Memory

D. Gopher N. Fussfeld W. Koenig D. Karis
Technion - ITT Univ. of Illinois
Haifa, Israel Champaign, Illinois

The paper describes the results of a sequence of experiments conducted on a two hand chord typewriter, to compare the efficiency of different coding principles employed to associate letters with their chord productions. This keyboard represents an effort to identify effective alternatives to the existing typewriter. It consists of two seperate 5-key panels (one for each hand), and letters are entered by typing chords composed of one to five fingers. Each panel is capable of producing the full Alphabet, and hence can be considered to constitute an independent typewriter. Two key questions raised by this design are: a) What is the best coding principle to represent the same letter on the two panels? b) Is there a relationship between the perceptual complexity of the chord patterns that represent different letters and the actual time to type them ?

One group of experiments contrasted the typing performance of three groups of subjects. Two groups were trained with horizontal, flat panels. The first acquired a code based upon hand symmetry. The second was given a spatial congruence coding principle based upon key arrangement. A third group was trained in a condition in which the panels were tilted upright to a vertical postion in which hand symmetry and spatial arrangements are unified in the coordination of the two hands, and all representation conflicts are resolved. The results show a clear superiority of the spatial over hand symmetry coding principle, and an additional advantage for the integration of the two in a vertical posture. Further experiments indicate that the main cause for these differences is the mode in which letter chords are represented in long term memory. If subjects are taght to conceive or imagine a superior representaion principle, their performance improves dramatically even if the actual performance conditions are conductive to high interference.

A second group of experiments was designed to separate between perceptual and motor factors in the acivation of single letter chords. The results underline the importance of perceptual factors in the activation of motor plans. The complexity of the patterns employed to represent letters was shown to account for 50 percent of variance in the typing speeds of single letters. Interesting individual differences were found in the reliance of individuals on visually based codes, and these differences affected systematicaly their typing performance.

The theoretical implications of these results are discussed in relation to a vision based theory of action plans.

Applied issues are raised with a reference to the design of data entry devices, and the training of psychomotor skills.

N85
14546

UNCLAS

N85 14546 D11

ON LOOKING INTO THE BLACK BOX:

PROSPECTS AND LIMITS IN THE SEARCH FOR MENTAL MODELS*

William B. Rouse and Nancy M. Morris

Search Technology, Inc.
25B Technology Park/Atlanta
Norcross, Georgia 30092

SUMMARY

It is a common assumption that humans have mental models of the systems with which they interact. In fact, it is difficult to explain many aspects of human behavior without resorting to a construct such as mental models. However, acceptance of the logical necessity of mental models can result in the raising of a whole new set of issues, many of which can be quite troublesome. For example, what form do mental models take? How does the form affect the usage of the models? How can and should a designer or trainer attempt to affect humans' mental models?

Despite many sweeping claims in the literature, available answers to the above questions are rather inadequate. There are prospects for improving this situation. However, there also

---

*This summary is based on the following report:
Rouse, W.B. and Morris, N.M. On looking into the black box: prospects and limits in the search for mental models. Norcross, GA: Search Technology, Inc., July 1984, 55 pp.

appear to be limits; the black box will never be completely transparent. This paper considers these prospects and limits.

To place the arguments advanced in this paper in perspective, alternative points of view with regard to mental models are first reviewed. Use of the construct in areas such as neural information processing, manual control, decision making, problem solving, and cognitive science are discussed. Also reviewed are several taxonomies of mental models.

Attention then shifts to the available empirical evidence for answering the questions posed earlier. A variety of studies are reviewed where the type and form of humans' knowledge have been manipulated. Also considered are numerous transfer of training studies whose results provide indirect evidence of the nature of mental models.

The alternative perspectives discussed above and the spectrum of empirical evidence are combined to suggest a framework within which research on mental models can be viewed. By considering interactions of dimensions of this framework, the most salient unanswered questions can be identified. Further, conjectures can be offered concerning possible inherent limitations in the search for mental models.

N85
14547

UNCLAS

N85 14547 D12

Issues in Developing a Normative Descriptive Model for Dyadic Decision Making

By

Daniel Serfaty

and

David L. Kleinman

CYBERLAB
University of Connecticut
Dept. of Electrical Engineering
and Computer Science
Storrs, CT 06268

ABSTRACT

Most research in modelling human information processing and Decision making has been devoted to the case of the single human operator. In this present effort, concepts from the fields of Organizational Behavior, Engineering Psychology, Team Theory and Mathematical Modelling are merged in an attempt to consider first the case of two cooperating decisionmakers (the Dyad) in a multi-task environment. Rooted in the well-known Dynamic Decision Model (DDM), our normative descriptive approach brings basic cognitive and psychophysical characteristics inherent to human behavior into a Team Theoretic analytic framework. An experimental paradigm, involving teams in dynamic decision making tasks, is designed to produce the data with which to build the theoretical model.

N85
14548

UNCLAS

LN85 14548 D13

Abstract(informal paper, Annual Mental)

GETTING MENTAL MODELS AND COMPUTER MODELS TO COOPERATE

Thomas B. Sheridan, James Roseborough, Leon Charney and Max Mendel

A qualitative theory of supervisory control is outlined wherein the mental models of one or more human operators are related to the knowledge representations within automatic controllers (observers, estimators) and operator decision aids (expert systems, advice-givers). Methods of quantifying knowledge and the calibration of one knowledge representation to another (human, computer, or "objective truth") are discussed. Ongoing experiments in the use of decision aids for exploring one's own objective function or exploring system constraints and control strategies are described.

PRECEDING PAGE BLANK NOT FILMED

Other Issues

N85
14549

UNCLAS

D14

N85 14549

A COMPARATIVE STUDY OF ALTERNATIVE CONTROLS AND DISPLAYS FOR BY THE SEVERELY PHYSICALLY HANDICAPPED*

Douglas Williams and Carol Simpson
Psycho-Linguistic Research, 2055 Sterling Av. Menlo Park, CA 94025
and Margaret Barker
Children's Hospital at Stanford, 520 Willow Road, Palo Alto, CA 94304

ABSTRACT

Effective communication for the physically disabled individual relies on control and display system design. The systems currently available to individuals restricted to using single switch interfaces by their involuntary (athetoid) movements have limited use and are inadequate for communication. Frustration followed by rejection of these aids is widespread.

An observed difficulty is the inability of a user to reliably activate a time-dependent control and display system. It is possible that this group of physically disabled individuals could control devices more adequately with an appropriately designed type of control and display system.

This study investigates a modification of a row/column scanning system in order to increase the speed and accuracy with which communication aids can be accessed with one or two switches. A selection algorithm was developed and programmed in BASIC to automatically select individuals with the characteristic difficulty in controlling time dependent control and display systems. Four systems were compared: 1) **Row/Column Directed Scan** (2 switches), 2) **Row/Column Auto Scan** (1 switch), 3) **Row Auto Scan** (1 switch), and 4) **Column Auto Scan** (1 switch). For this sample population, there were no significant differences among systems for scan time to select the correct target. **The Row/Column Auto Scan system resulted in significantly more errors** than any of the other three systems. Thus, the most widely prescribed system for severely physically disabled individuals turns out for this group to have a higher error rate and no faster communication rate than three other systems that have been considered inappropriate for this group.

BACKGROUND

This project addressed a question that was raised during the development of a versatile, portable, speech prosthesis (VPSP) for the severely disabled (LeBlanc, Simpson, Williams, and Lingel, 1980), which is a

---

* A joint effort between the Rehabilitation Engineering Center, Children's Hospital at Stanford and PLRA, supported by an Office of Special Education, U.S. Dept. of Education grant. Maurice LeBlanc, Director of Research, Childrens Hospital at Stanford, directed the project. The software was developed by Sol Katzman. Research assistants Becky Gordon, Tom Dominguez and Leslie Roberts helped run the subjects.

microprocessor based, wheelchair portable, speech prosthesis that can be controlled to speak and/or store any speech message desired by the user. Different control and display systems were provided for users with varying degrees of motor control: 1) 1-switch, row/column scanning; 2) joystick + selection switch or 5-slot controller and a user-driven cursor; 3) keyboard, direct selection. These control and display systems were chosen because they are typical of a wide range of commercially available communication and environmental control devices.

While communication speeds of 30 words per minute were obtained using the VPSP keyboard, experienced users could do no better than 4 WPM using the row-column scanning system, and users with cerebral palsy were considerably slower. Wethered (1976) found that individuals with cerebral palsy were slower using single switches than individuals with muscular dystrophy or multiple sclerosis, supporting what was found with the VPSP.

Commercially available communication and environmental control aids all use variations of three basic approaches to control and display systems which enable the user to indicate his or her intended elements or symbols of communication. These are scanning, encoding and direct selection. The approach that an individual uses is dependent upon that individual's physical and cognitive abilities. For an individual physically limited to the use of one or two switches or methods of indication, the approach used must involve scanning, although it may be combined with encoding in some cases. (Vanderheiden and Harris-Vanderheiden, 1976).

Vanderheiden (in Vanderheiden and Grilley, 1976) indicated that scanning is extremely powerful because it can be used by individuals with minimal control (e.g. able to consistently make only one or two movements or signals). He also indicated that the power of the scanning technique is offset by slow speed of communication. The speed is slow because much time is spent passing over unwanted symbols before arriving at the desired symbol.

During the VPSP user evaluation, it was observed that individuals with athetoid cerebral palsy and involuntary movements could not reliably activate the single switch at the desired item, regardless of cursor speed (as slow as 6 sec. per jump). No measurable improvement was found in a week of continuous use. Speech therapists reported similar problems with other devices using the row-column scanning system; yet devices with that control system continue to be prescribed non-vocal individuals with involuntary movements. And, apparently taking their cue from existing devices, designers of new, microprocessor based systems for the handicapped are using the row/column scanning system, for example, Bruey, 1980.

APPROACH

Our purpose was to study the ability of persons with athetoid cerebral palsy to control a two-switch, user-driven cursor, row/column scanning system. The underlying assumptions made were 1) That "1-switch" users will actually be able to control some other, second switch. (This assumption is supported by the observation that all users of the 1-switch VPSP were able to deal with two switches, the on/off switch and the selection switch); and 2) That those with severe athetosis who are 1-switch users would, as a group, be able to perform a regular, successive switch activation. Some observations of VPSP operators using "verify on" mode led us to suspect this would be true. (LeBlanc, et. al, 1980)

An interface consisting of two switches utilized this behavior. One switch, the "scanning switch", moved the cursor when it was actuated (speed of cursor movement was adjustable for each user). This switch was used to position the cursor on the desired item - one of the rows or an item within a row. The second switch was the "selection switch", which "selects" the item next to the cursor. If the selected item is a row to be scanned, then the act of selecting it causes the cursor to move across the row when the user next activates the scanning switch. If the selected item is an item in one of the rows, then the act of selecting it transfers the selected item to a holding space (where the user would construct sentences for viewing, printing or speaking in a complete prosthesis system.) When neither switch is pressed, the system does nothing, allowing the user to rest, think, or whatever.

Our principal goal was to determine whether the alternative user-driven cursor, 2-switch system is any more effective for the target population than is the currently used row/column scanning system. A one-dimensional vertical column linear scan system, and a one-dimensional, linear (horizontal row) scan system. were also studied so as to isolate the effect, if any, of the two-dimensional feature of row/column scanning systems apart from their scanning feature.

METHOD

Candidate Subjects

From CHS patient lists and from students at local schools, we obtained volunteers for this study and from this group chose six for intensive study who had: 1. severe athetosis; 2. cognitive ability to recognize the letters used and to follow the directions given; and 3. passed the screening test (described below).

Screening Test

The purpose of the screening test was to select only those individuals who are neither extremely accurate nor totally random in their operation of a row/column scanning display and control system. We wanted to include in the study only those individuals who were like the VPSP users that we observed erroneously selecting the item just before or just after the one they wanted. We refer to the volunteers who took the screening test as "candidates". Those who were selected by the screening test (and the other 3 criteria above) became the subjects for the study.

The screening test consisted of from 3 to 6 trials, with the number of trials determined by the candidate's performance as the test progressed. The task for each trial was to use the "select" switch to select a target item, a letter of the alphabet, presented by the computer. Only the first response on each trial was recorded for test purposes. Each wrong response was recorded in terms of its distance in the scan in front of or after the target: -1 means the item in the scan just before the target; 0 means the target was accurately selected; +3 means the item 3 items after the target was selected. The probability of someone whose switch-actuation behavior is completely random selecting one particular letter from an array of 12 letters is 0.0833. Thus, the chance of three such selections being made in three trials by such an individual is 0.00057. This was our criterion for accurate performance; if an individual selected the target letter three times, he was judged not impaired enough to qualify for this study. If he made three +1 or -1 errors (p = 0.0046), then the candidate was judged to have the requisite athetoid behavior and comprehension of the task to participate in the study. If, after three trials, neither 3 target selections nor 3 +1 or -1 errors had occurred, the testing was continued to a maximum of six trials.

If at the end of 6 trials, 3 or more of the responses were of the -1, 0, or +1 type, with the condition that 2 of the 3 be of the -1 or +1 type (p = 0.0033), the candidate was included as a subject in the study. If any other pattern of responses resulted, the candidate was not selected, since a variety of other problems, such as inability to find the target visually, could be present. (Such problems are also important but are simply beyond the scope of this study.)

Subjects

Of the 39 candidate subjects who participated in the screening process, 18 were too accurate, 15 displayed neither high accuracy nor the high rate of +1 and -1 type of errors. The remaining 6 candidates were selected by the algorithm as having the behavior of interest to this study. Table 1 shows the characteristics of the six subjects in terms of age, diagnosis, and current communication device, if any.

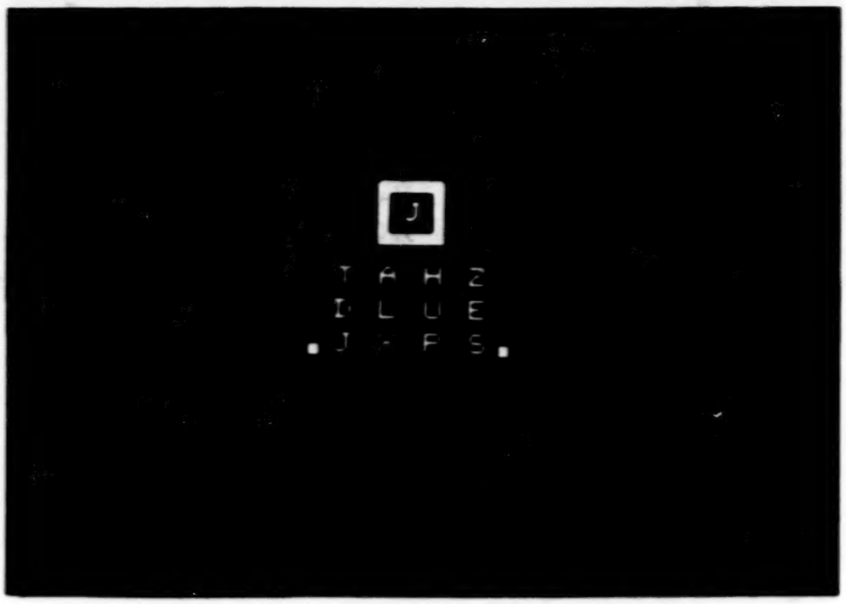
Equipment

A TRS-80 Model III Microcomputer was purchased for this project, and programmed to handle all stimulus presentation, response timing and accuracy measures, and some of the data reduction. Suitable single and double switches were constructed by the Children's Hospital at Stanford Rehabilitation Engineering Center to provide each child with a switch he or she could activate as reliably and accurately as possible, by whatever method was most effective. Actuation means included head motion, head stick, hands, feet, or gross limb movement. The computer was programmed to present an easily-taught and understood letter selection task on the CRT screen. Twelve letters of the alphabet were chosen for their visual dissimilarity, as determined from grapheme confusion matrix data in the published literature (Kinney, Marsetta, and Showman, 1966). These letters were displayed on the screen in a row/column matrix, a single row, or a single column, as appropriate for the experimental condition.

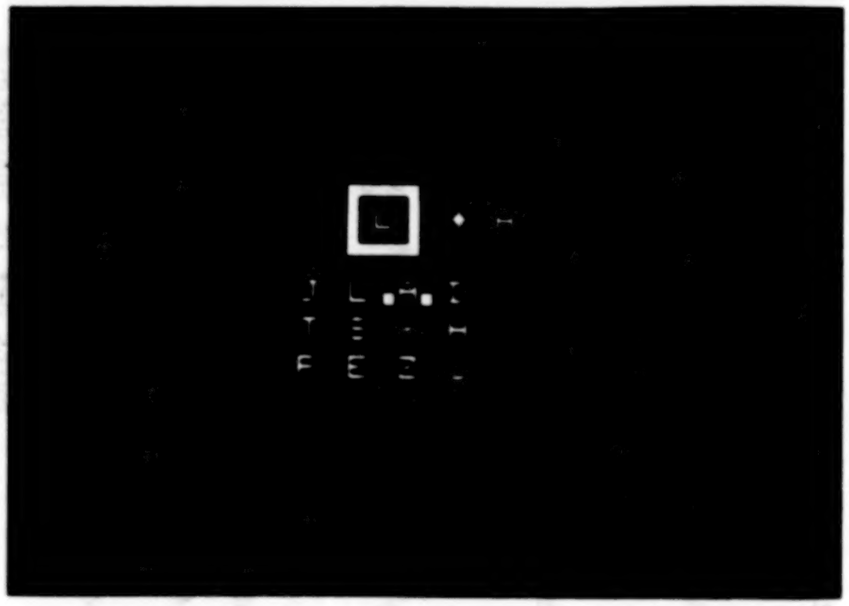
Subj. #, Sex	Age	Diagnosis	Current Communication Method	Elec.	Selection	Scanning Control
1,F	14	Athetoid CP	ETRAN and eyes left and right for yes and no answers to questions	none	Zygo lever left side of head	Zygo lever right side of head
2,F	18	Athetoid CP	Speech unintelligible except to friends; no written method	none	Right hand pressing Zygo tread switch on table	Left hand pressing Zygo tread switch on table
3,M	14	Athetoid CP	Some speech, nods for yes/no, some typing with many errors	joy-stick on elec. w/c	Zygo tread switch on laptray (L. hand)	Zygo tread switch on laptray (R. hand)
4,F	20	Athetoid CP	Nods head for yes/no; eyegaze at wordbook	none	Zygo tread switch on L. side of headrest	Zygo tread switch on R. side of headrest
5,M	18	Athetoid CP Severe Scoliosis	Yes/no by looking at chair arms; "yes" and "no" written on w/c arms	none	Zygo tread switch on L. side of headrest	Zygo tread switch on R. side of headrest
6,M	18	Athetoid CP	Speech unintelligible except to friends; Blissboard, with difficulty due to inaccurate pointing	joy-stick on elec. w/c	L. hand on Zygo tread switch on laptray	R. hand on Zygo tread switch on laptray

TABLE 1.
CHARACTERISTICS OF SUBJECTS

ORIGINAL PAGE IS
OF POOR QUALITY



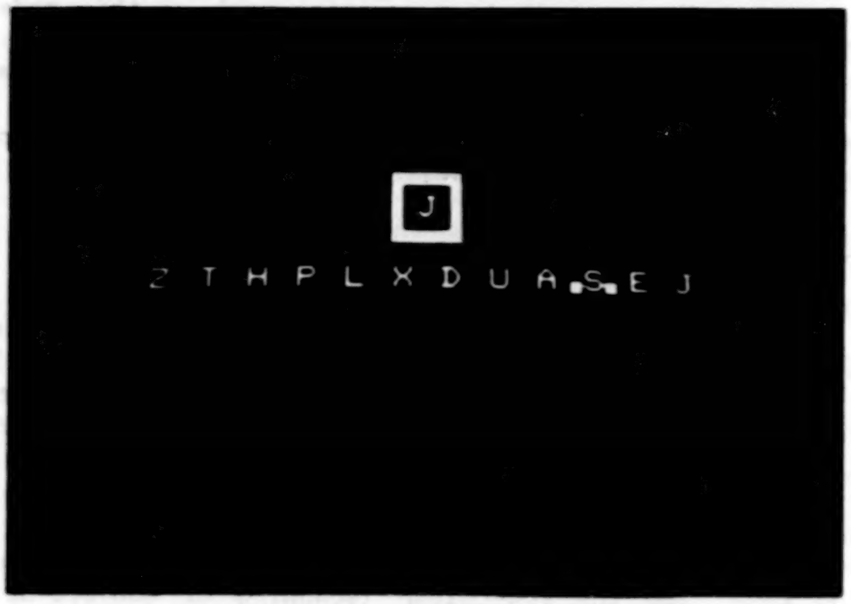
a. Row column scanning system, row indicated by cursors.



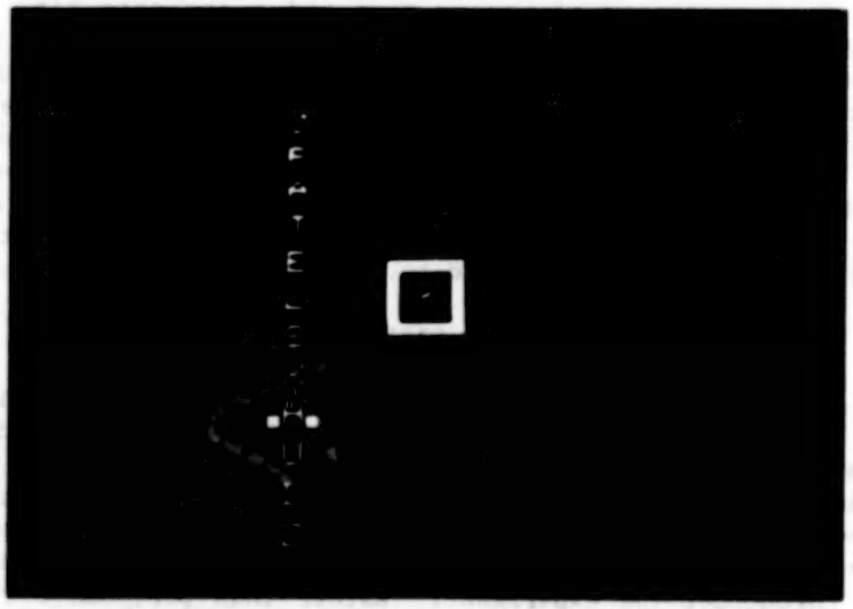
b. User driven cursor, 2-switch system. Letter "a" indicated by cursors.

Figure 1. Photographs of displays on the TRS-80 Mode III Monitor for each of the four types of systems. The "target" letter is displayed in the box.

ORIGINAL PAGE IS
OF POOR QUALITY


J
Z T H P L X D U A S E J

c. Row scanning. Letter "s" indicated by cursors.



d. Column scanning. Letter "h" indicated by cursors.

Figure 1, continued.

(See Figure 1) The computer was programmed to drive the cursor in a row-column scan under computer control, a row- column scan under subject control, and two types of linear scan (vertical and horizontal) under computer control. The programs to accomplish much of this made use of algorithms implemented in software developed for the VPSP. Implementation of these algorithms to provide a test and evaluation package on a widely-available computer is a significant benefit of this project, and this software has been made available for diagnosis of children and evaluation of switches.

Experimental Design

Each of the six subjects used each of the four systems, one at a time. The order of systems used was balanced across subjects to control for transfer effects. A given subject used a system approximately every other day, except weekends, for approximately 2 hours per day. The first day consisted of familiarization with the system to reduce the effects of learning. Each subsequent day with that system included some refamiliarization time. Each day contained 3 runs of 12 trials each. Between runs, subjects rested and relaxed, as they wished. The task for each trial was to select a "target" letter (which was shown on a special area of the CRT) from the row-column (scanning or subject-driven) or linear array displayed on the CRT. Each of the twelve letters was presented an equal number of times within each run, in random order. A subject used the same system for 3 days for a total of 9 runs with that system. The procedure was then repeated with the next system until that subject had used all four systems. All runs with a given system were in succession to provide for subject familiarization with a given system and so that learning could be detected.

Data collected consisted of errors made and a response time measure called scan time. Scan time was defined as the cumulative time that the cursor was actually actively scanning in the case of the auto scan systems or the time that the cursor was actively controllable by the subject, in the case of the directed scan system. All responses (both erroneous and correct) were recorded. After 6 erroneous responses for any single trial, however, the computer terminated the trial automatically. We assumed that further attempts by the subject to select the correct response would only lead to fatigue and frustration.

RESULTS AND DATA ANALYSIS

Two problems made the results of the experiment more difficult to analyze. First, due to procedural problems, some runs were not done with the system which had been chosen for that run. Examination of the data showed that the interpolated wrong runs did not affect the times. Thus, the problem with order was ignored in all data analyses. The second problem was that not all subjects were able to use the same machine-paced scan rate throughout the experiment. Rather than postpone a run until the subject was able to operate the switch at a previously-run rate, runs were completed at a rate which the subject could use. Thus, subjects 2,3, and 6 used two different rates. Their times to successfully select the target letter are affected on those runs where they used the different rate.

From the point of view of pure science, this is a problem with the data because it confounds the variables of scan time and system. However, from a practical point of view, one could argue that on that day, that particular person could not have operated at the faster rate (or would have made many more errors in so doing) so the longer times are realistic reflections of his capability with that particular system.

Scan Time Data Analysis

Table 2 shows the overall means and standard deviations of scan time for each of the four systems.

TABLE 2. MEAN SCAN TIME SCAN TIME IN SECONDS FOR FOUR SYSTEMS ACROSS 9 RUNS PER SYSTEM AND 6 SUBJECTS.

SYSTEM	SCAN TIME MEAN	SD
ROW/COL DIRECTED	19.4	6.9
ROW/COL AUTO	15.1	6.5
ROW AUTO	15.0	5.0
COLUMN AUTO	15.8	8.1

Table 3 shows the results of a 2-Way AOV of the effects on scan time for systems by runs for all six subjects. (AOV done with computer programs from Wollach, 1983)

TABLE 3. ANALYSIS OF VARIANCE FOR SCAN TIMES

SOURCE	SS	DF	MS	F	p
Subjects	5947.7	5	1189.5	26.82	<0.0001
System	693.2	3	231.1	1.72	0.2
Subj. x Sys	2005.0	15	133.7		
Runs	783.2	8	97.9	2.06	0.063
Runs x Ss	1902.9	40	47.6		
Sys x Runs	1356.6	24	56.5	1.76	0.024
Sys x R x Ss	3852.6	120	32.1		
Total	16541.3	215			

There was no significant effect for system ($p=0.2$), but runs approached significance ($p=0.063$). The interaction between systems and runs was significant ($p=0.024$) however, as was subjects ($p<.0001$).

Error Data and Analysis

Table 4 shows the total number of misses made by all subjects for each of the four systems.

TABLE 4. TOTAL NUMBER OF MISSES MADE BY SIX SUBJECTS FOR EACH OF FOUR SYSTEMS.

	SYSTEM			
	ROW/COL DIRECT	ROW/COL AUTO	ROW AUTO	COLUMN AUTO
TOTAL MISSES	249	451	289	281

The error analysis, Table 5, is a simple 1-way AOV with system as the variable under test, and a significant (p=0.0005) effect due to system is shown.

TABLE 5. ANALYSIS OF VARIANCE TABLE FOR ERROR DATA

SOURCE	SS	DF	MS	F	p
System	49.3	3	16.4	6.18	<0.001
Subjects	319.9	5	64.0	24.10	<0.001
Sys x Ss	107.5	15	7.2	2.70	<0.005
Within	504.9	190	2.6		
Total	981.6	213			

Duncan's Multiple Range Test (Bruning & Kintz, 1968) shows that the only system significantly different from any of the others is row-column auto scan, significantly worse than any of the others.

Table 6 shows the total number of successful target selections and failures in target selection for each system. For a given target, the selection was scored as a success as long as the subject managed to select the target correctly within six tries. If he or she missed six times on the same target, then that trial was scored as a failure and a new target was presented.

TABLE 6. NUMBER OF FAILURES AND OF SUCCESSFUL TARGET SELECTIONS FOR EACH OF FOUR SYSTEMS. N=6 SUBJECTS, EACH PERFORMING 108 TARGET TRIALS PER SYSTEM.

	ROW/COL DIRECT	ROW/COL AUTO	ROW AUTO	COLUMN AUTO
FAILURES	2	13	2	7
SUCCESSES	646	635	646	629*

A Chi-Square analysis showed a significant effect for system ($X^2=13.77$, df=3, p=.0032). As can be seen from the table above, the Row/Column Auto Scan system's failure rate was nearly double that of the next worse system - Column Auto Scan - and was over six times the rate of either the Row/Column Direct or the Row Auto Scan system.

CONCLUSIONS

The significant subjects effect was expected, even after the subjects had passed thru the rigorous selection procedure. It is the nature of these subjects that their performance is variable and idiosyncratic. What this study re-emphasized is that good scientific research can be done on these people, and that one does not have to resort to "clinical judgment" when dealing with a diverse population like this.

Learning seems to have been eliminated as a factor in the results. There was a slight improvement from run one to run nine irrespective of the system used; most of this gain occurred in the first three runs. The "runs" factor was only marginally significant, and the counterbalancing prevented learning from affecting any one system differentially; so we can conclude that learning or practice was probably not a factor in the differences found between systems.

An important implication of this last finding is that the two hours practice was sufficient time for these subjects to have reached at least a learning plateau, if not an asymptote, in the use of these systems. While these were fairly simple control/display systems the subjects did have to adapt to new switches, a new situation, people they had not met before, a novel task, and a computer which they had not worked with before. It is useful to know that two hours practice is sufficient to almost completely train this population. This finding alone could save many hours of unnecessary "training" which people like these are often subjected to when they are introduced to a new communication system.

---

*Procedural problems due to fatigue of one subject reduced the total number of trials for the Column Auto Scan system to 636 instead of the 648 trials used for the other three systems.

Examination of the mean scan times shows that the **row-column auto-scanning** scheme is, for these subjects, the **worst** choice, though not significantly so, of all the systems tested. In terms of **errors**, it is **significantly worse** than any other tested system.

It should be noted here that these subjects were persons who would normally be prescriptively limited to a single-switch input device, since their motor control is so poor. It is a very important finding of this study that such people can in fact run a two-switch directed-scan system, and experience less frustration due to fewer errors, and in addition can equal or better the output speed of the three auto scan types of systems which are traditionally prescribed for them. We emphasize that we did not look for persons who could perhaps operate two switches; rather, the subjects who were selected by our screening program later demonstrated an ability to use two switches. This represents a doubling of channel capacity in the information theory sense. Certainly all attempts should be made for this group of users to find a second switch that they can reliably operate when not under the rigid time pressure of an automatic scanning system.

It seems that there is something about the arrangement of items to be selected into rows and columns which makes it more difficult for these people to deal with them. A simple analysis of scan rate and the number of items that must, on the average be passed over in order to reach the target, would seem to dictate putting items into rows and columns in order to minimize scan time. Information theory supports this traditional approach, telling us that square matrices are more efficient than those having unequal numbers of rows versus columns. For users who can use a scanning system with high accuracy this might be so, but it is certainly not the case for the subjects included in this study. We are fairly confident in recommending that people who meet the selection criteria outlined earlier in this report should not be given a row-column autoscan system as a communication aid. Our data show that instead, a row autoscan system would work well for them if enough could be fit onto a single row to meet their needs; if not, they should be trained to use two switches and fitted with a row-column directed scan system. We would predict that their error rate would be lowest with the row-column directed scan, and their speed would be close to that achieved with the single row. A major contribution of this project is, in fact, the ease with which persons in this group can be identified. Anyone with a Radio Shack Model I or Model III computer can perform this diagnostic test merely by requesting a copy of the program from Children's Hospital and having the potential user run the subject selection program.

Readers who are interested in a fuller treatment of this study may request "A Comparative Study of Control and Display Design Principles Which Affect Efficient Use of Communication Aids by the Severely Physically Disabled-- Final Report" from Children's Hospital at Stanford, 520 Willow Road, Palo Alto, CA 94304. The software used is available on 5" disklettes from the same source.

References

Barker, Margaret R. and Albert M. Cook, 1980. Matching Device Characteristics to client needs and system performance measures. **Proceedings of VOCA Conference,** May, 1980, Berkeley, CA. (in press).

Bruning, J. L., and Kintz, B. L. 1968. **Computational Handbook of Statistics,** Scott, Foresman, & Co.

Bruey, A. J., 1980. Microcomputer Hardware for the Handicapped: Single-key data entry for the PET, **Kilobaud Microcomputing,** November, 1980, p 173.

Kinney, G.C., Marsetta, M., and Showman, D.J., 1966. Studies in display symbol legibility, part XXI. The legibility of alphanumeric symbols for digitalized television. Bedford, Mass: the Mitre Corporation, ESD-TR-66-117. (as cited in Lindsay, P.H. and Norman D.A., **An Introduction to Psychology,** New York: Academic Press, 1972, 122-23.

LeBlanc, M.A., Simpson, C.A., Williams, D.H., Lingel, C.D., 1980. **Progress Report - Research and Development of a Versatile Portable Speech Prosthesis,** NASA Ames Grant No. NSG-2313, Childrens Hospital at Stanford, 520 Willow Rd., Palo Alto, CA 94304.

Montgomery, J., 1980. Measuring effectiveness of communication aids with children and adults, Talk presented to the Meeting of the Bay Area Non-Oral Communication Group, May 13, 1980, Mauzy School, Alamo, California.

Vanderheiden, G. C., 1981. Technically Speaking, **Communication Outlook,3,**2:10

Vanderheiden, G.C., 1976. Providing the child with a means to indicate, **in Vanderheiden and Grilley, eds., Non-Vocal Communication Techniques and Aids for the Severely Physically Handicapped,** Trace Center, University of Wisconsin, pp. 20 ff.

Wethered, Chris E., 1976. **Hierarchies of Operational Efficency and Preference for Interface Selection for the Physically Disabled.** Masters Thesis, University of Idaho.

Winer, B. J., 1962. **Statistical Principles in Experimental Design,** McGraw-Hill Book Co., Inc.

Wolach, A. H., 1983. **BASIC Analysis of Variance Programs for Microcomputers,** Brooks/Cole Publishing Co., Monterey, California.

Zygo Industries, Inc. pahphlet, **An Improved Row/Column Scanning System; Tetrascan II,** 1983, Zygo Industries Inc., P.O. Box

N85
14550

UNCLAS

D15

N85 14550

A MANUAL CONTROL TEST FOR THE DETECTION AND DETERRENCE OF IMPAIRED DRIVERS

Anthony C. Stein
R. Wade Allen
Henry R. Jex

Systems Technology, Inc.
13766 S. Hawthorne Blvd.
Hawthorne, CA 90250

Telephone No. 213/679-2281

ABSTRACT

A brief manual control test and decision strategy have been developed, laboratory tested, and field validated which provide a means for detecting human operator impairment from alcohol or other drugs. The test requires the operator to stabilize progressively unstable controlled element dynamics. Control theory and experimental data verify that the human operator's control ability on this task is constrained by basic cybernetic characteristics, and that task performance is reliably affected by impairment effects on these characteristics.

Assessment of human operator control ability is determined by a statistically based decision strategy. The operator is allowed several chances to exceed a preset pass criterion. Procedures are described for setting the pass criterion based on individual ability and a desired unimpaired failure rate. These procedures were field tested with apparatus installed in automobiles that was designed to discourage drunk drivers from operating their vehicles. This test program, sponsored by the U.S. Department of Transportation, demonstrated that the control task and detection strategy could be applied in a practical setting to screen human operators for impairment in their basic cybernetic skills.

INTRODUCTION

This paper reviews the development and validation of a behavioral testing device which can detect human operator impairment. These skills are important in performing tasks which require continuously manipulating displayed variables with a control device, such as driving or machinery operation. The manual control skills required to perform these types of tasks have been extensively studied (Ref. 1), and the test described herein has been developed to detect impairment in these skills.

The test involves two components, a control task and a detection strategy. The control task, called the Critical Tracking Task (CTT), was developed in the early sixties to test pilot and astronaut visual motor performance (Refs. 2 and 3). Over the years it has proven to be an effective indicator of the effects of environmental stresses [e.g., noise (Ref. 4); space station

PRECEDING PAGE BLANK NOT FILMED

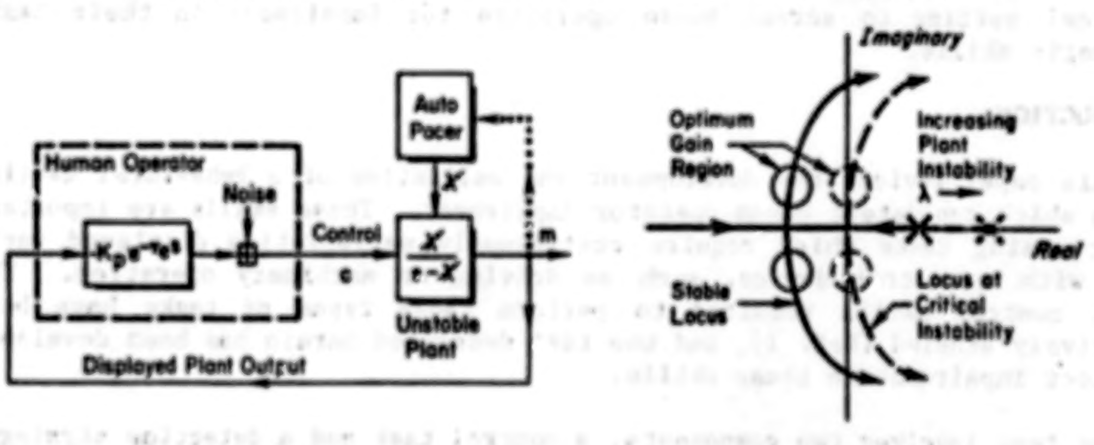
confinement (Ref. 5); ship motion, (Ref. 6); spacecraft re-entry (Ref. 7); human operator impairment (e.g., alcohol (Ref. 8); and marihuana (Ref. 9)].

The use of the CTT as an alcohol impairment detection device was first tested in automobiles by the General Motors Corp. (Ref. 10). Subsequent research sponsored by the U. S. Dept. of Transportation (Refs. 11-14) was conducted to optimize the test strategy. In subsequent research the statistical decision theory for optimizing the detection strategy was developed and validated in laboratory tests (Ref. 15). Following this, vehicle mounted devices were assigned to convicted drunk drivers to obtain field validation data (Ref. 16).

In the remainder of this paper we will briefly describe the control theory basis for the CTT, the statistical theory behind the impairment detection strategy, and laboratory and field test results which validate tester performance in a practical, operational environment.

Critical Tracking Task (CTT)

The task description and theory of operation for the CTT have been previously documented (Refs. 2 and 3). A summary is illustrated in Fig. 1. The task dynamics consist of an unstable controlled element, and an autopacer unit which controls the location of the unstable pole. No input is necessary because the operator's remnant (noise) is sufficient to disturb the system. The unstable root, λ, is initially set at a small value. As the subject begins to perform the task, the plant instability is increased (the root moves further into the right half plane). When a filtered version (with a one second time constant) of the displayed plant output deviations (m) exceed about 15% of the display range, the rate of increase of λ is reduced by a factor of four times in order to slowly approach the point of closed loop instability and avoid overshoot. When m exceeds the display limits, control loss is assumed and the pole location at this point, termed the critical instability limit or λ_c, is used as the task performance metric. The total test time for experienced subjects is on the order of 30 seconds.


Auto Pacer
Human Operator
Noise
$K_p e^{-\tau_e s}$
Control
c
λ
$\frac{\lambda}{s-\lambda}$
m
Unstable Plant
Displayed Plant Output
Imaginary
Optimum Gain Region
Increasing Plant Instability
$\lambda \rightarrow$
Real
Stable Locus
Locus at Critical Instability

a) Block Diagram

b) Root Locus

Figure 1. CTT Task Elements and Root Locus Stability Analysis

As indicated in Fig. 1, the subject's task performance depends on a visual/motor dynamic time delay (τ_e), a gain (K_p), and internal noise or remnant (random control actions). The subject's time delay dictates the shape of the root locus (the pure time delay causes the complex branches to bend to the right) while K_p determines the operating point on the locus. Increasing the task instability (λ) translates the entire locus to the right or unstable direction. The pure gain closure dictates two primary close loop roots (a pure time delay actually gives an infinite number of roots, but the lowest frequency pair dictate the stability characteristics). The operator's optimum strategy is to set K_p to locate both closed loop poles on the imaginary axis as indicated. The task is continually perturbed by the operator's internal noise source. As the point of closed loop instability is approached, the underdamped closed loop system response tends to increasingly amplify display deflections, at first causing a reduction in the autopacing rate, then finally terminating a trial when the display bounds are exceeded. These theoretical aspects have been carefully validated by experiments in the USA (Refs. 2 and 3); and Netherlands (Ref. 17).

Impairments can affect the human operator's control capability in three ways: 1) increased visual/motor time delay (τ_e); 2) interference with accurate K_p adjustments; 3) increased noise. Any combination of these three impairment effects would tend to reduce the achievable task score, λ_c. Several past studies have been conducted on the effects of alcohol on λ_c. Summary results are plotted in Fig. 2. As noted here, results have been extremely reliable across several past studies.

Impairment Detection Strategy (IDS)

Details of the development and optimization of the IDS have been described previously (Refs. 15 and 18). The objective of the IDS is to maximize the chance of detecting operator impairment with a minimum number of CTT trials. This research developed a statistically based decision strategy to maximize test discriminability (i.e., low fail rate for normal operators and high

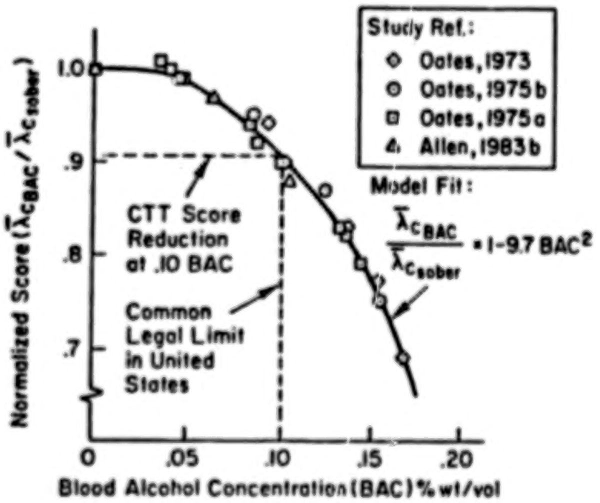
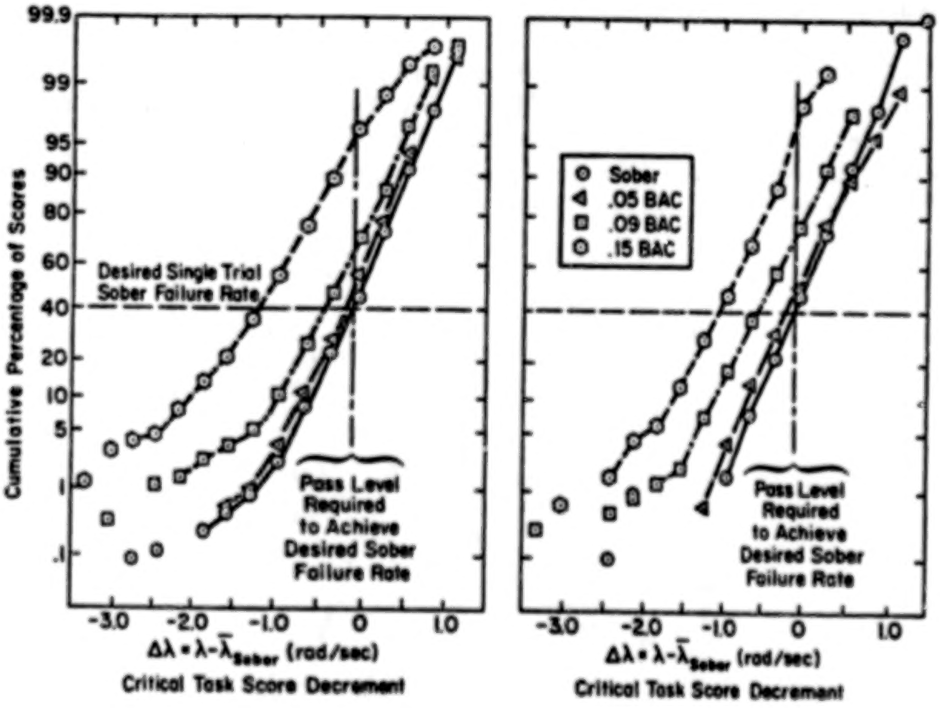


Figure 2. Experimental Results of Alcohol Effects on CTT Performance Over Several Past Research Studies

failure rate for impaired operators). The IDS development and optimization started with an analysis of the statistical properties of CTT performance (λ_c). Analysis of past data showed trial-to-trial and between subject performance variability to be quite consistent across several studies (Refs. 15 and 18) and a reliable effect of alcohol impairment was noted as illustrated in Fig. 2. It was also found that subjects could be rapidly trained on the CTT but residual long term skill improvement would have to be accounted for.

Based on the statistical analysis of past data several IDS requirements were established: 1) significant performance differences between operators require individualized pass criteria; 2) stable performance score variance and relatively independent trial-to-trial performance variability allow the use of simple multiple sampling strategies; and 3) long term residual skill improvement would require procedures for sampling and periodically upgrading performance criteria.

The important statistical characteristics of CTT performance relative to IDS development can be illustrated with cumulative distribution functions as shown in Fig. 3. The distributions are normalized and averaged across a large number of subjects and plotted on probability paper (a Gaussian distribution plots as a straight line). The data are normally distributed over a wide range, and the alcohol effect is clearly indicated. The basic requirement of the IDS is that sampled subject performance must exceed a preset pass level. Several sampling strategies were analyzed and tested with past data bases (Refs. 15 and 18) and, for various reasons, a one pass out of several



a) Data From Oates, 1975b

b) Data From Oates, 1975a

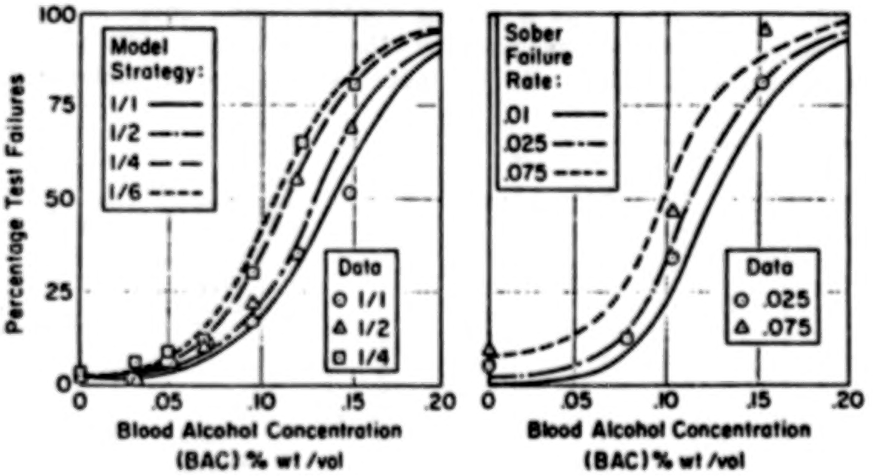
Figure 3. CTT Differential Score Distributions Averaged Across 24 Subjects in Each Experiment

permitted attempts was selected. With this strategy, and assuming independent trials, the probability of failing the test is the single trial probability of failure raised to the power of the number of permitted attempts:

$$P_{fail}(N \text{ trials}) = [P_{fail}(\text{single trial})]^N$$

This approach permits us to simply define the pass level given a subject's performance distribution and a desired probability of test failure when sober. For example, for a 2.5% failure probability given four attempts, the single trial probability must be approximately 40% (i.e., $(0.4)^4 \approx 0.025$). Given this sober pass level as indicated in Fig. 3, one can also derive the expected drunk failure rates (i.e., at BAC = 0.10, $(0.76)^4 \approx 0.35$ and at BAC = 0.15, $(0.96)^4 \approx 0.85$). A statistical model based on the above procedure was developed, and IDS model predictions of failure rates were compared with failure rates obtained with the IDS applied to past experimental data (Ref. 15). The discriminability results are illustrated in Fig. 4.

The good agreement above between model and data suggest that the detection strategy is well understood, and that an adequate procedure for establishing task performance pass levels has been established. The above strategy and procedures embody two other desirable features: first, the pass levels are near a subjects average or median performance level, which is stable and can be determined reliably; second, a subject's cumulative distribution function can be used to easily determine pass level, and also to upgrade the pass level to account for residual skill improvement.



a) Variation in Permitted Attempts at Constant Sober Failure Rate = .025

b) Variation in Sober Failure Rate for a One Pass in Four Attempts (1/4) Strategy

Figure 4. Impairment Detection Strategy Comparison Between Model Analysis and Experimental Data (Strategies Involve One Pass in Several Attempts — 1/N)

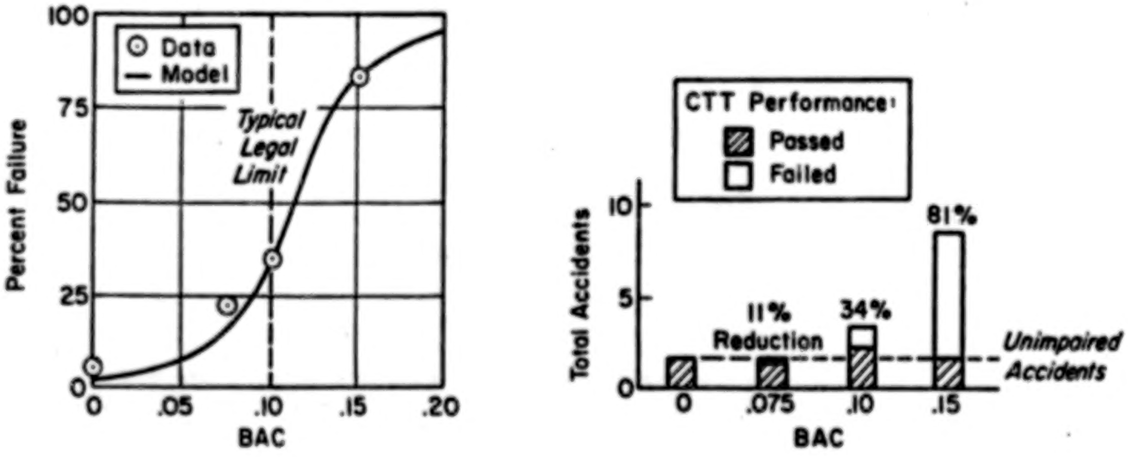
LABORATORY VALIDATION EXPERIMENT

To validate the effectiveness of the CTT and IDS just described, an experiment was conducted that compared CTT score with both BAC (Blood Alcohol Concentration; weight/volume) and driving performance in a driving simulator (Ref. 15). Subjects were convicted drunk drivers obtained through the cooperation of the Los Angeles Municipal Courts. Twenty-four so called volunteers were permitted to participate in the experiment as a condition of probation, and, in exchange, received a reduction in their court sanctioned fine.

After being accepted, the subjects were required to participate in three 2-hour training sessions and three full-day experimental sessions. Each subject participated in one placebo and two drinking sessions. The subject population was divided into three groups matched for age, sex, and driving experience, and the order of occurrence of placebo session was different for each group.

Validation experiment results are summarized in Fig. 5. Notice, first, that the discriminability data agree with the statistical model developed from past experimental studies. More importantly, analysis of simulator data shows a high correlation between simulator accidents and test failure. As shown in Fig. 5, pre-drive CTT failures detected 81% of subsequent simulator accidents. These correlations between predicted and actual test performance show that it is now possible to both predict and verify vehicle operator impairment using a cybernetic task such as the Critical Tracking Test in combination with a suitable Impairment Detection Strategy.

Additional findings were also obtained on subject training procedures. CTT performance obviously has a strong motivational component. The validation experiment subjects were assigned by the traffic court and were not truly motivated volunteers. Several subjects exhibited a lackadaisical attitude



a) CTT Impairment Discriminability, Data and Statistical Model Comparison

b) CTT Failure Detection of Simulator Accidents

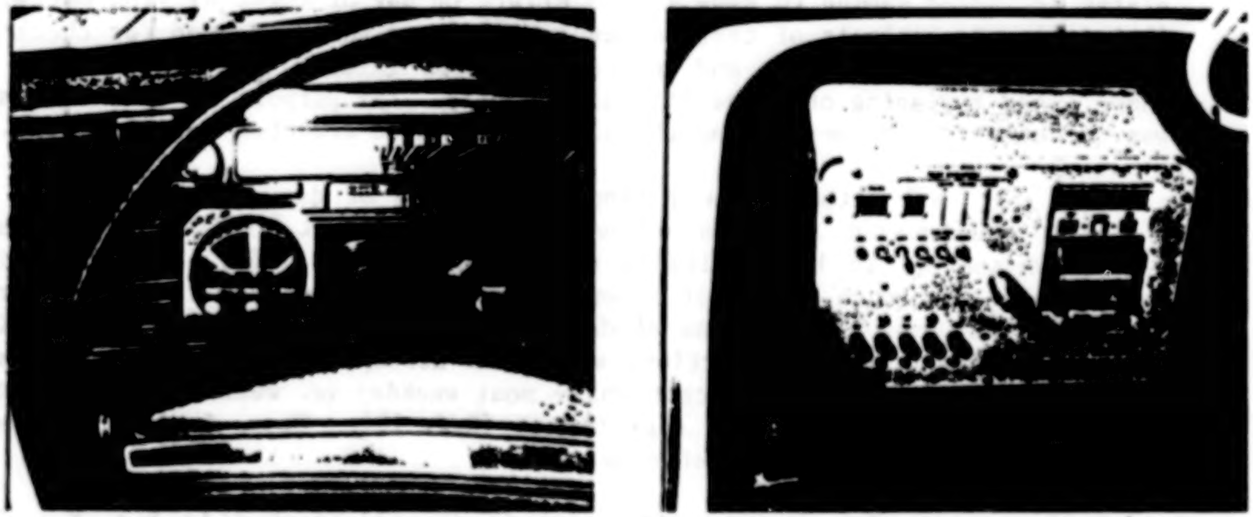
Figure 5. Results from Laboratory Validation Experiment. One Pass in Four Attempts Detection Strategy, Desired Sober Failure Rate = 0.025

during training, and were not encouraged by the positive reinforcement payments that were offered for good performance. In a subsequent training experiment (Ref. 19) it was found that giving a time penalty (30 second wait) for test failures was a much more effective way to deal with non-volunteer subjects who were motivated mainly to minimize their time involvement.

FIELD VALIDATION EXPERIMENT

The purpose of the field validation experiment was to demonstrate that a vehicle mounted CTT/IDS could be assigned to convicted drunk drivers on a practical basis. This included selection and assignment by traffic courts and exclusive routine use by the recipients for a six month period. The vehicle mounted test equipment, shown in Fig. 6, combined the CTT/IDS into a system called the Drunk Driving Warning System (DDWS); and was installed in ten 1978 Chevrolet Nova autos. The subject had to pass the DDWS test in order to deactivate certain alarms: the car could be driven without passing the test, but in this case the emergency flashers would operate, and if the car was driven over 10 miles per hour, the horn would honk once per second. If the driver failed the test (four fail trials in succession), the computer required a ten minute wait before retesting was permitted.

Various countermeasures were incorporated into the DDWS to prevent cheating. These included sealing components and cables to prevent or reveal physical tampering, and requiring retesting if the driver left the driver's seat or opened his door after starting the test. An event recorder was also incorporated into DDWS to monitor the driver's use of DDWS and record instances of test failure and/or driving with alarms activated. Extensive usage data by time of day were obtained.



a) Subject Display and Steering Wheel Control

b) Trunck Mounted Electronics and Cassette Data Recorder

Figure 6. Vehicle Mounted Field Test Apparatus

Two municipal court judges were willing to administer the DDWS as a sanction to convicted drunk drivers. The California law was temporarily modified to permit experimental evaluation of the DDWS sanction. Approval and/or cooperation was obtained from various state agencies (e.g., the Department of Motor Vehicles) in order to carry out the field test program. Nineteen drivers were subsequently assigned DDWS vehicles to be used exclusively over a six month period. Their licenses were restricted so that they could not legally drive any other vehicle. After initial training the alarm system was activated and the subjects were required to check-in at two week intervals. During the check-in sessions, the car and DDWS system were inspected, and the data tape was removed and computer analyzed. The subjects were debriefed and questioned about test failure episodes. (There was no penalty for admitting such instances during the test period.)

RESULTS

The overall results were derived from three basic data sources: 1) recorded data which was retrieved and reviewed at the biweekly check-in sessions; 2) in-depth assessments developed during data reviews and debriefings at the biweekly check-ins; 3) structured interview data obtained from subjects, colleagues, and relatives of subjects, court and state agency personnel associated with the program, and others associated with the general drunk driving problem. Results from these three sources were as follows.

Recorded Data

Recorded data were analyzed to look for DDWS influence on driving patterns, subject performance, and the ability of DDWS to detect impaired drivers. Requiring the driver to take the CTT test with or without the DDWS alarms activated seemed to have little effect on day or night driving patterns (Ref. 16). An analysis of test passes and failures was performed for check-in sessions at the beginning and end of Phase II (alarms on) and the end of Phase I and beginning of Phase III (alarms off). The purpose of this analysis was to determine whether having the alarms activated affected vehicle usage.

Data for test attempts as a function of time of day are illustrated in Fig. 7. Chi-squared analysis showed the test attempt differences between alarms on and off to be marginally significant ($p = 0.038$). On a relative basis the alarms on vs. off test attempts are small except for the early morning hours (12:00-4:00 am). Time of day differences were obviously highly significant. Time of day interactions with test attempts and performance (pass/fail) were found to be significant while most weekday vs. weekend interactions were found to be small or not significant (Ref. 18). Thus, further analysis was restricted to time of day effects.

Failure rates for various time periods are illustrated in Fig. 8. Day time failure rates were about what was expected (i.e., $\approx$ 2.5 percent) based on the procedure used to set individualized CTT pass scores. Nighttime failure rates were three to seven times greater than this level, however, which is consistent with high incidence of drinking and driving during nighttime "recreational/social" periods vs. daytime trips for commuting to and from work and running errands.

ORIGINAL PAGE IS
OF POOR QUALITY

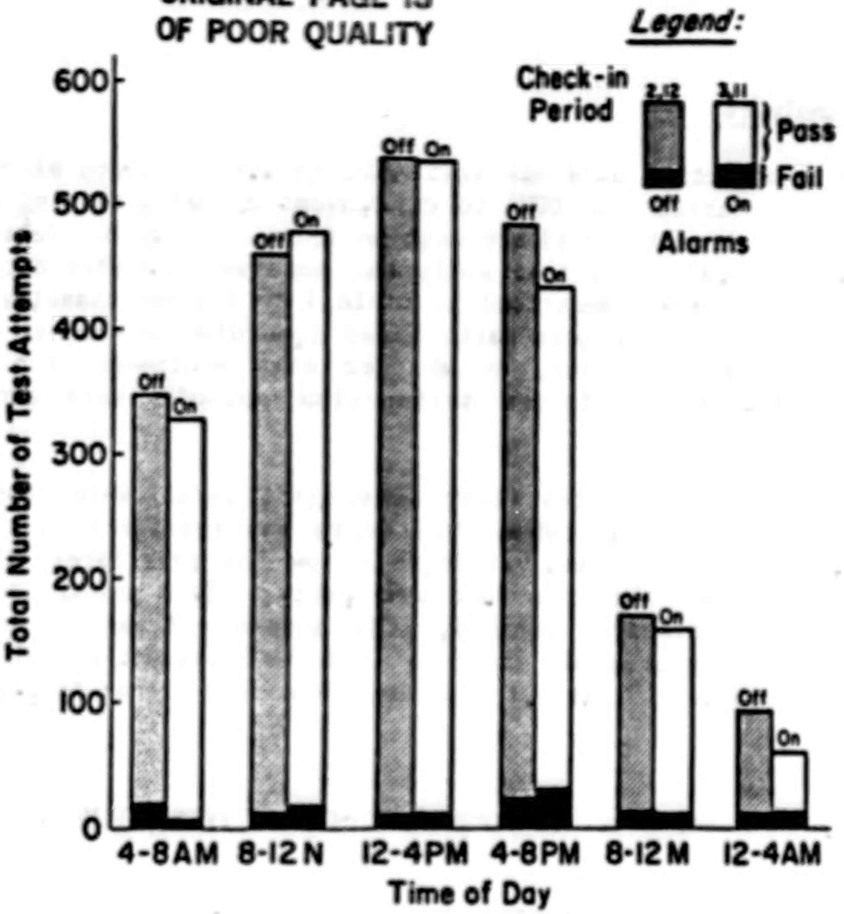


Figure 7. Effects of Alarms on Test Attempts and Performance (Pass/Fail) as a Function of Time of Day

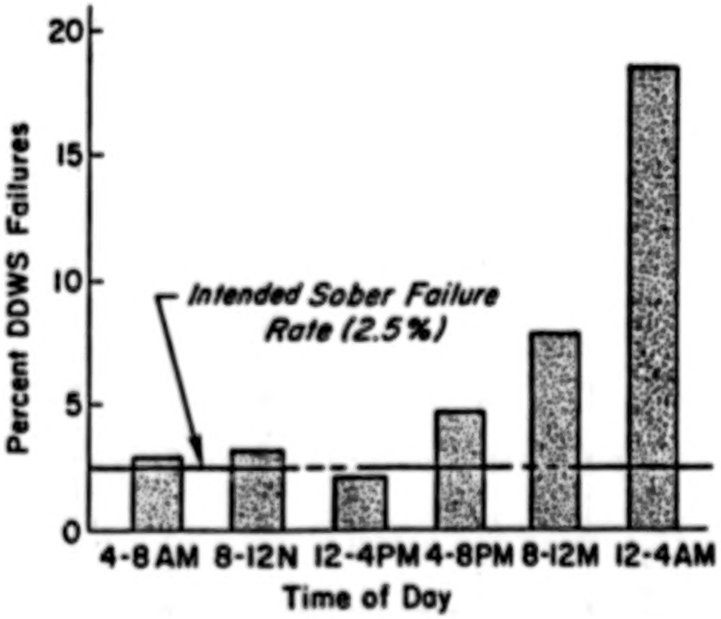


Figure 8. DDWS Failure Rate as a Function of Time of Day

In-Depth Analysis

Since no objective data was available on subject blood alcohol concentration (BAC), the ability of DDWS to circumvent drinking/driving trips rests on circumstantial objective evidence such as shown in Fig. 8. Debriefing information was obtained on all test failures, however, and this data was combined with objective data as summarized in Table 1 to further classify test failure. Total test failures have been partitioned according to whether the driver was felt to be sober, impaired, or whether some equipment problem might have caused the failure (equipment malfunction episodes were experienced with several subjects).

Differential test scores (test score—pass level) were computed from the cassette logged data, and when this score was greater than −0.4 (i.e., the test score was greater than 0.4 units below the pass level) the subject was assumed to be sober when the test was taken. This assumption was based on analysis of a statistical model for CTT scores and amounts to a 95 percent level of confidence that BAC was less than 0.05 percent wt/vol (Ref. 18). In the case of subject 19, it was felt that his pass level in the beginning was

TABLE 1. IN-DEPTH ANALYSIS OF TEST FAILURES

Subject Number	Test Failures				
	Total	Sober ($\Delta\lambda_p > -0.4$)	Problem	Impaired	Trips With Alarms
01	36	9	3	24	0
05	20	6	8	6	0
06	5	3	0	2	0
07	4	2	1	1	0
08	6	4	1	1	0
09	4	3	1	0	0
10	14	9	2	3	0
11	8	4	1	3	0
12	17	9	1	7	0
13	38	26	6	6	0
14	29	12	12	5	0
15	6	5	0	1	0
16	4	3	2	0	0
17	26	5	10	11	1
19	112	24*	8	81	5
20	13	12	0	1	1
22	9	4	2	3	0

*$\Delta\lambda_p > -0.2$

set too high, so his total failures for $\Delta\lambda > -0.2$ were used. Problem failures were interpreted from the in-depth analysis, and the impaired failures were given by

$$F(\text{Impaired}) = F(\text{Total}) - F(\text{Sober}) - F(\text{Problem})$$

As noted in Table 1, even if we account for sober and problem failures, there still remain a significant portion of impaired failures, with two subjects accounting for the majority of these. The DDWS alarms should deter the subject from driving, but as recorded by the data logger and indicated in Table 1 three subjects drove with the alarms activated.

Subjects 17 and 20 had isolated incidences where the car had to be moved a short distance. Subject 19 actually admitted to occasionally driving his car without passing the test after drinking. This constituted a fairly serious violation of one of the conditions of probation, and the court was so notified. Subject 19 was cooperative, however, and it was recommended that he be permitted to remain in the program.

Debriefing Information Analysis

The municipal courts and California Department of Motor Vehicles carried out their part in project support without serious problems. The courts do need an individual to take charge of subject screening, however, as was available through the West Los Angeles Municipal Court. Also, license restriction needs to be indicated on the front of the license to alert enforcement personnel and others (e.g., car rental agencies) of the DDWS user's restricted driving privilege. California is currently investigating this feature and may provide it in the near future.

Public acceptability for the DDWS concept has been quite good, once the objectives, approach, and background have been fairly presented. News media accounts of DDWS were fair and many times positive, although occasionally with some minor misinformation. Positive opinions have also been elicited by other individuals associated with the drunk driving problems, including relatives and colleagues of the DWI offenders employed here as subjects.

Finally, subject acceptance was quite good. No one found the DDWS to be a hardship, and most found it to be a desirable and effective sanction. Most subjects would choose DDWS compared to fines, license restriction or suspension, or jail.

CONCLUDING REMARKS

The data presented here and elsewhere (Ref. 18) indicate that a DDWS equipped vehicle can maintain good impaired driver discriminability in a field setting. As to whether subjects drive after test failure, in-depth analysis showed only three subjects drove with the alarms on (a violation of probation which is recorded by the DDWS data logger). One subject was determined to have driven while impaired, and even in this case there is some indication that the drive was made at low speed. Thus, test failure would appear to significantly deter DWI trips.

The CTT/IDS could be used as a cybernetic screening device in other scenarios such as: daily screening of commercial or government vehicle operators, industrial process or power plant operators, etc. "Card/key" systems could be used to permit a common device to be used by a number of individuals wherein the individual scores are updated in the card via a magnetic strip. Finally, the impairment detection system (IDS) could be used with other cybernetic tasks that might prove to be sensitive to other aspects of human operator impairment.

REFERENCES

1. McRuer, D. T., and E. S. Krendel, Mathematical Models of Human Pilot Behavior, AGARDograph No. 188, Jan. 1974.

2. Jex, H. R., J. D. McDonnell, and A. V. Phatak, A "Critical" Tracking Task for Man-Machine Research Related to the Operator's Effective Delay Time. Part I: Theory and Experiments with a First-Order Divergent Controlled Element, NASA CR-616, Nov. 1966.

3. McDonnell, J. D., and H. R. Jex, A "Critical" Tracking Task for Man-Machine Research Related to the Operator's Effective Delay Time. Part II: Experimental Effects of System Input Spectra, Control Stick Stiffness, and Controlled Element Order, NASA CR-674, Jan. 1967.

4. Allen, R. Wade, Raymond E. Magdaleno, and Henry R. Jex, Effects of Wide Band Auditory Noise on Manual Control Performance and Dynamic Response, AMRL-TR-75-65, Oct. 1975.

5. Allen, R. Wade, and Henry R. Jex, Visual Motor Response of Crewmen During a Simulated 90-Day Space Mission as Measured by the Critical Task Battery, NASA CR-2240, May 1973.

6. Jex, Henry R., Richard J. DiMarco, and Warren F. Clement, Effects of Simulated Surface Effect Ship Motions on Crew Habitability — Phase II. Vol. 3: Visual-Motor Tasks and Subjective Evaluations, Systems Technology, Inc., TR-1070-3, Feb. 1976.

7. Jex, Henry R., Richard A. Peters, Richard J. DiMarco, and R. Wade Allen, The Effects of Bedrest on Crew Performance During Simulated Shuttle Reentry. Vol. II: Control Task Performance, NASA CR-2367, Oct. 1974.

8. Klein, Richard H., and Henry R. Jex, "Effects of Alcohol on a Critical Tracking Task," J. Stud. Alcohol, Vol. 36, No. 1, 1975, pp. 11-20.

9. Stoller, Kenneth, George D. Swanson, and J. Weldon Bellville, "Effects on Visual Tracking of Δ^9-Tetrahydrocannabinol and Pentorbarbital," J. Clinical Pharmacology, Vol. 16, 1976, pp. 5-6.

10. Tennant, Jean A., and Richard R. Thompson, "A Critical Tracking Task as an Alcohol Interlock System," SAE Paper 730095, Jan. 1973.

11. Sussman, E. D., and C. N. Abernethy, Laboratory Evaluation of Alcohol Safety Interlock Systems. Vol. I, DOT, 1973 (available from NTIS).

12. Oates, John F., Jr., Experimental Evaluation of Second-Generation Alcohol Safety-Interlock Systems, DOT-TSC-NHTSA-73-9, 1973.

13. Oates, John F., David F. Preusser, and Richard D. Blomberg, Laboratory Testing of Alcohol Safety Interlock Systems, Phase II, Dunlap and Associates, Darien, Conn, 1975.

14. Oates, J. F., Jr., D. F. Preusser, and R. D. Blomberg, Laboratory Testing of Alcohol Safety Interlock Systems. Vol. I: Procedures and Preliminary Analyses, Dunlap and Associates, Inc., 1975.

15. Allen, R. Wade, Anthony C. Stein, and Henry R. Jex, "Detecting Human Operator Impairment with a Psychomotor Task," Proc. of the 17th Annual Conf. on Manual Control, JPL Publ. 81-95, 1981, pp. 611-625.

16. Allen, R. Wade, Anthony C. Stein, Leland, G. Summers, et al., Drunk Driving Warning System (DDWS). Vol. II: Field Test Evaluation, Systems Technology, Inc., TR-1136-1-II, Dec. 1983.

17. Stassen, H. G., "Application of Describing Function Methods," in Progress Report — January 1970 Until January 1973 of the Man-Machine Systems Group, Delft University of Technology, Report WTHD 55, 1973.

18. Allen, R. Wade, Anthony C. Stein, Leland G. Summers, et al., Drunk Driving Warning System (DDWS). Vol. I: System Concept and Description, Systems Technology, Inc., TR-1136-1-I, Nov. 1983.

19. Cook, Marcia, Henry R. Jex, Anthony C. Stein, et al., "Using Rewards and Penalties to Obtain Desired Subject Performance," Systems Technology, Inc., P-288, Proc. of the 17th Annual Conference on Manual Control, University of California, Los Angeles, CA, June 1981; also JPL Publ. 81-95, 15 Oct. 1981, pp. 211-222.

N85
14551

UNCLAS

ↄ/16

N85 14551

ELECTROMYOGRAPHIC PATTERNS ASSOCIATED WITH DISCRETE LIMB MOVEMENTS

D. M. Corcos, G. L. Gottlieb, and G. C. Agarwal

Department of Physiology, Rush Medical College, Chicago, and
Department of Industrial and Systems Engineering,
University of Illinois at Chicago

ABSTRACT

The relationship between the movement time (MT) for accurate and rapid discrete movements of distance A to a target of width W was quantified by Fitts and is given by the equation:

$$MT = a + b \log_2 (2A/W)$$

This relationship, known as Fitts' Law, has received considerable support for many types of movements. It also raises the interesting question: if MT is affected by distance moved and accuracy, then how do the patterns of muscle activation alter?

Recent studies on elbow joint movements indicate that for movements of different amplitudes, either the intensity of the EMG or the time course increases with increasing distance.

We studied how accuracy of movement affects the patterns of muscle activation. The study was performed on the ankle joint because of the asymmetrical nature of extensors and flexors. Seven subjects made accurate and rapid ankle movements of 12, 18 and 24 degrees to targets of 2, 4 and 8 degrees. The data suggest that the agonist muscle was activated for a longer time and with greater intensity for larger movements. The duration of the EMG burst increases for increases in target size but the amplitude was not affected. It appears that the pattern of activation is modified in bot intensity and duration according to task demands.

Data is presented to show the effect of adopting different movement strategies on the pattern of muscle activation and the consequent velocity profile. The interrelationship of various kinematic and EMG variables is considered.

PRECEDING PAGE BLANK NOT FILMED

INTRODUCTION

The neurophysiological mechanisms underlying voluntary, purposeful, rapid limb movements from one position to another has been a fertile topic of research for centuries. Descartes (1637) proposed an elaborate control model of agonist and antagonist activity.

One typical pattern of muscle activation associated with self terminating movements has been described as triphasic. This refers to the fact that in electromyographic activity (EMG) there is an agonist burst, an antagonist burst and a second agonist burst. In a general sense this pattern seems most compatible with an impulse-timing theory of movement control (Schmidt, Zelaznik, Hawkins, Frank and Quinn, 1979; Wallace, 1981; Wallace and Wright, 1982) in which movement is controlled in terms of successive bursts of activity that propel the limb and then arrest the movement. From this theory, predictions have been generated relating different parts of the EMG pattern to different types of movement.

In its simplest form the impulse-timing model suggests that the agonist burst propels the limb and the antagonist burst decelerates the limb. This theory is similar to the bang-bang control of an intertial load using minimum time criterion (Smith, 1962). Neither theory requires a third burst.

If movements are regulated in this way, then the following questions arise. When a limb is moved successively further distances, what are the controlled parameters of the EMG? What are the effects of asking a person to voluntarily change the speed of his movements? Many of these factors involve trade-offs and the literature is not clear on these issues.

If we consider the mechanism for moving different distances Freund and Budingen (1978) suggest that target directed movements require approximately the same time no matter how large the distance covered. This finding is very important since it implies a mechanism which adjusts velocity to hold time constant, a finding in contradiction with Fitts' Law (1954, 1964) that predicts that movement time should systematically increase as movement distance is increased or target size is decreased. Close inspection of the procedures used by Freund and Budingen, however, suggests that their data is compatible because they allowed required target accuracy to decrease as movement distance and velocity increased, the well-known speed-accuracy trade-off. Ghez (1979) has also provided evidence in favor of keeping duration relatively constant and postulates a mechanism for modulating peak acceleration. Neither article reports a systematic evaluation of the EMGs. Lestienne (1979) does and also points out that movement distance does not affect agonist duration. This finding also receives support from Brown and Cooke (1981) who maintain that "Although the graph in Fig. 3B suggests a trend for this burst to increase in duration as movement amplitude increased, regression analysis indicates that any such change was non-significant." (p. 101).

In direct contrast to the above findings, Wadman, Denier van der Gon, Geuze and Mol (1979) suggest that EMG amplitude stays constant and duration increases as distance moved increases. Also, Angel (1974) found that the duration of the initial EMG volley is prolonged. Enoka (1983) has extended the speed control system hypothesis to multidirectional and multiarticular movements and has shown that both skill level and direction of movement have an effect on the duration of the net flexor torque. In these experiments, however, neither movement velocity nor movement time were controlled.

How might these findings be reconciled? One possibility is that velocity is a regulated factor and is controlled by a mechanism which adjusts the rate of rise of tension in the muscle. As long as movements are generated at less than maximal velocity, moving faster is accomplished by increasing the rate of tension rise (increasing EMG amplitude) and holding duration constant. However, this system is rate limited as is obvious from Hill's equation (1938) for the force-velocity relationship. The same mechanism could be used to move a limb further. However, when a large distance needs to be covered as rapidly as possible, duration may have to increase if maximum rate of rise of tension cannot occur in the time taken for the short movement. Duration would also have to increase if the maximum rate of rise of tension in a given duration is insufficient for the required movement.

As a starting point, let us assume that Fitts' Law does hold for the types of movement being considered. Fitts' Law relates movement time (MT) to distance moved and target size in the following manner: MT = a + b log2 (2A/W), where a and b are empirically determined constants, A = distance moved and W = the target width. With this in mind, let us the consider the implications of the pulse-step model for the first agonist burst. According to Ghez (1979), rapid limb displacements of the cat are controlled by a pulsatile output which is modulated in amplitude and of approximately constant duration. This suggests that an equation for agonist duration (AD) would be reduced to a constant which would be some measure of the time for muscle activation (see Freund, 1983, pp. 420-421). If distance moved requires a change in the time of the first burst, then a closer approximation is AD = a + c A, where c is the scaling factor. If both distance and target size are important, a formula for agonist duration could resemble AD = a + c (A/W). It should be obvious that this is indeed a simplification since agonist intensity is also altering as a function of distance and ideally this should be incorporated. However, this will require a better understanding of the relationship between EMG and force in non-isometric movements (Agarwal and Gottlieb, 1982).

Studies which quantitatively relate the first antagonist burst to various movement requirements are few. As Hallet and Marsden (1979) have pointed out, antagonist EMG is notoriously unreliable and numerous authors have failed to present it. One of the distinguishing factors of the triphasic pattern is that the antagonist burst occurs during the silent period between the two agonist bursts; i.e., the bursts occur successively and not concurrently. However, evidence is accumulating that movements are not necessarily controlled by alternating bursts. As early as 1776, Winslow (see Tilney and Pike, 1925) suggested coactivation was important in movement control. Corcos (1982) found approximately 20 percent of all trials were triphasic and that coactivation occurred at least as frequently when individuals made rapid movements to a target.

Smith (1981) has suggested that alternating patterns (which may or may not be triphasic) occur when:

1) resistance prevents displacement,

2) in rhythmic movements, e.g., locomotion or mastication, or
3) movements are of low velocity or low muscular tension.

Coactivation occurs when:

1) tension needs to be precisely controlled or
2) movements are made at high velocity.

The relationship of the antagonist burst to velocity has been studied by Lestienne (1979). He suggested its purpose is to act as a brake when the agonist force exceeds the passive viscoelastic tension developed by the extensor and flexor muscles. For high speed movements, there was a substantial overlapping in agonists and antagonists during the acceleration component.

Mustard and Lee (1983) also analyzed the braking function. They showed that as movement amplitude increased, the excitation levels of the muscles increased but duration remained constant. They also considered the effect of movement velocity. A well defined antagonist burst did not occur until the velocity exceeded 250 degrees per second. With further increases in velocity, the antagonist burst increased in size and occurred earlier in the course of the movement. In apparent contrast, Brown and Cooke (1981) maintain that antagonist duration increases as distance covered increases but EMG activity remains constant. Waters and Strick (1981) showed that it was abolished when it showed no functional purpose.

Another candidate for regulation has been proposed by Hogan (1984). His suggestion is that the antagonist modulates mechanical impedance to maintain posture. In a simple demonstration, he showed that there are significant levels of simultaneous activation involved in maintaining upright arm posture and that the level of coactivation increases as gravitational torque increases. It would, therefore, seem plausible that coactivation would increase as torque, caused by factors other than gravity, increases. Hogan's study is one of the few in the area that considers that during contraction, muscles (both agonists and antagonists) are not merely force generators but variable compliances as well. This compliant property is probably of major importance in understanding how movements terminate but this issue has been overlooked.

No studies have been found which relate movement accuracy to the underlying control messages portrayed by surface EMG. That is, recent neurophysiological studies which use EMG have ignored much of the relevant psychological literature such as that of Fitts, who did not measure EMG.

Therefore, the purpose of this paper is to consider how patterns of muscle activation are altered by: 1) moving different distances, 2) moving to targets of different sizes. Human subjects were asked to make rapid flexion and extension movements about the ankle and elbow joints. The elbow joint has been studied by numerous investigators. The ankle joint is included in this study because of very significant differences in biomechanical properties of the flexor and extensor muscles and their respective role in posture.

METHODS

All experiments were performed using seven normal human subjects. A subject sat comfortably in a chair with his/her foot placed on a footplate so that the medial malleolus was aligned with the axis of rotation. When a comfortable position was established, the foot was tightly secured by velcro straps.

A video monitor was positioned about a meter in front of the subject so that different combinations of movement conditions could be displayed. The video monitor display was generated by an Apple II microcomputer. The Apple generated a display of the starting position for a movement, the amplitude of the movement and the width of the target. A signal corresponding to limb position was also displayed. A general purpose computer (SPC-16, General Automation) was used to control the experiment and record data on four input channels: the joint angle, the torque, and the rectified and filtered EMGs for the flexor and extensor muscles. The foot plate is attached to a torque motor with position, velocity and torque feedback to simulate different load dynamics. A very similar apparatus is used for elbow movements.

For the ankle joint, the EMGs were recorded from the tibialis anterior (flexor) and the soleus muscles (extensor). For the elbow joint, the EMGs were recorded from the biceps (flexor) and the triceps muscles (extensor). The details of EMG processing are given in Agarwal and Gottlieb (1977).

RESULTS

The data reported here represents preliminary observations on seven subjects. The subjects were asked to dorsiflex the ankle joint 12, 18, and 24 degrees to targets of 2, 4, and 8 degrees width. They were asked to make accurate movements as fast as possible. They also made 18 degree plantar-flexion movements to targets of 2, 4, and 8 degrees width.

Figure 1 shows an individual making 12, 18 and 24 degree movements to a 2 degree target as well as a movement with no target constraint. It can be seen that the movements are generated by discrete bursts of activity which are scaled to the movement requirements. Scaling occurs in both the agonist and antagonist muscles.

Figure 2 shows the same individual making an 18 degree plantar flexion movement to a 2 degree target. If EMG1 is considered, it can be seen that considerably more activation occurs when soleus is acting as the agonist than when it is acting as the antagonist (compare to EMG1 in BKA108 from Figure 1). The same finding is true for EMG2, tibialis anterior.

Figure 3 shows another subject peforming the same set of movements. If we compare the condition in which there was no target constraint to the conditions in which there were, it can easily be seen that there are substantial differences in the kinematics. The EMG patterns also reveal concomitant changes. The movements to the targets are very slow and precise and are regulated by a continuous signal which cannot be partitioned into discrete bursts.

Figure 4 portrays an 18 degree plantar flexion movement to a 2 degree

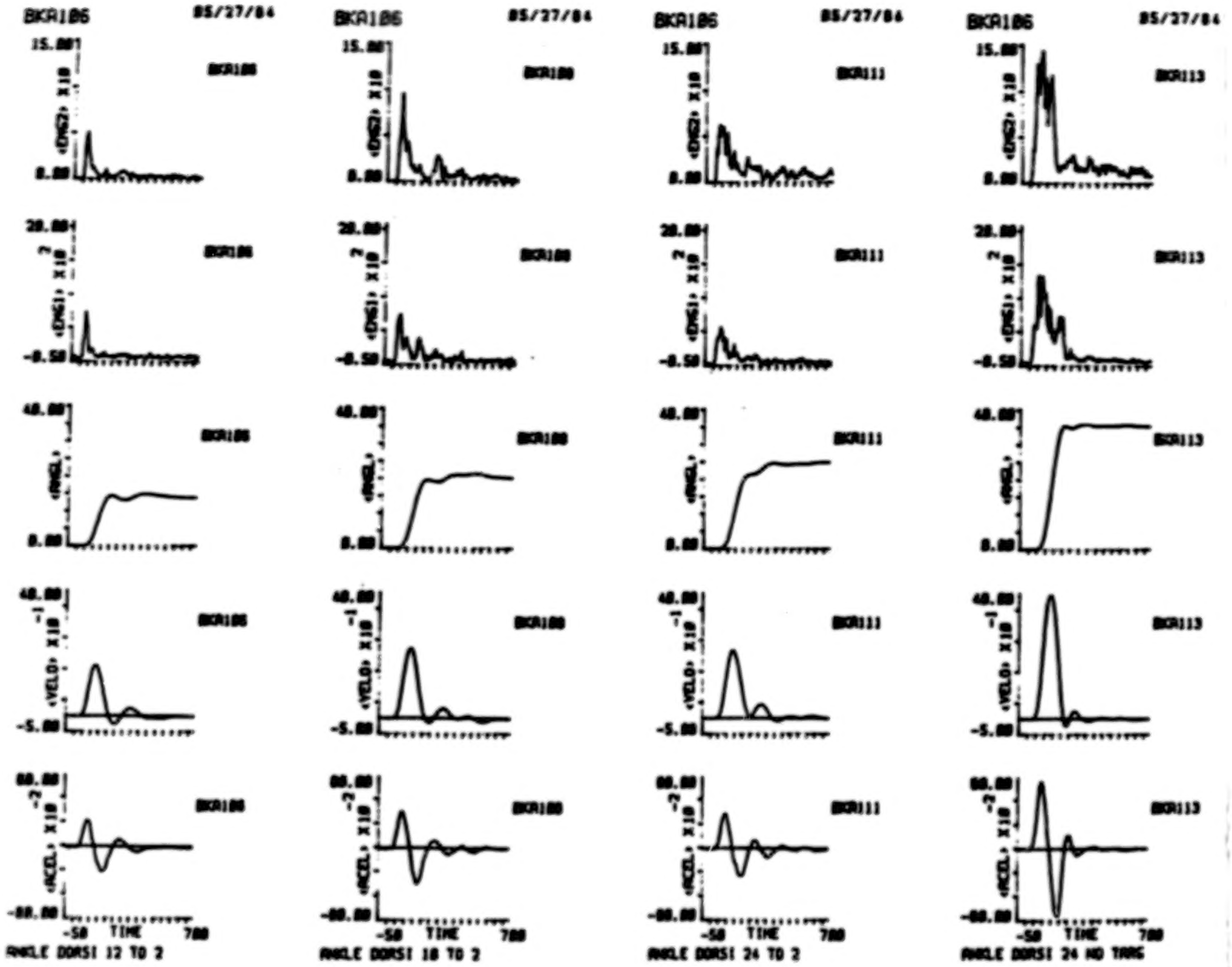


Figure 1 shows average data (approximately 10 trials) for a subject making 12, 18 and 24 degree movements to a two degree target (BKA106, BKA108, BAK111). BKA113 depicts a 24 degree movement to no target. EMG2 refers to tibialis anterior which is the agonist. EMG1 refers to soleus which is the antagonist. Angle data is in degrees, velocity in degrees per second and acceleration in degrees per second per second.

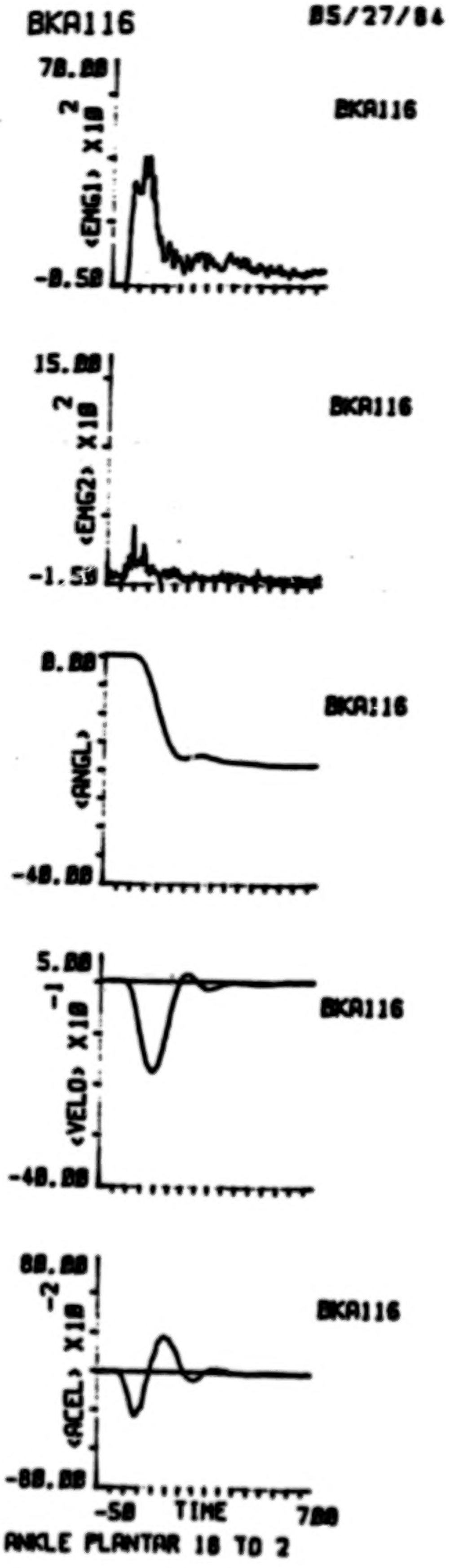


Figure 2 shows average data from the same subject as in Figure 1. The individual is making an 18 degree plantarflexion movement to a 2 degree target. EMG1 refers to soleus which is now acting as the antagonist. It should be noted that it's scale is 3.5 times larger than that in Figure 1. EMG2 refers to tibialis anterior whose scale is one tenth of that in Figure 1.

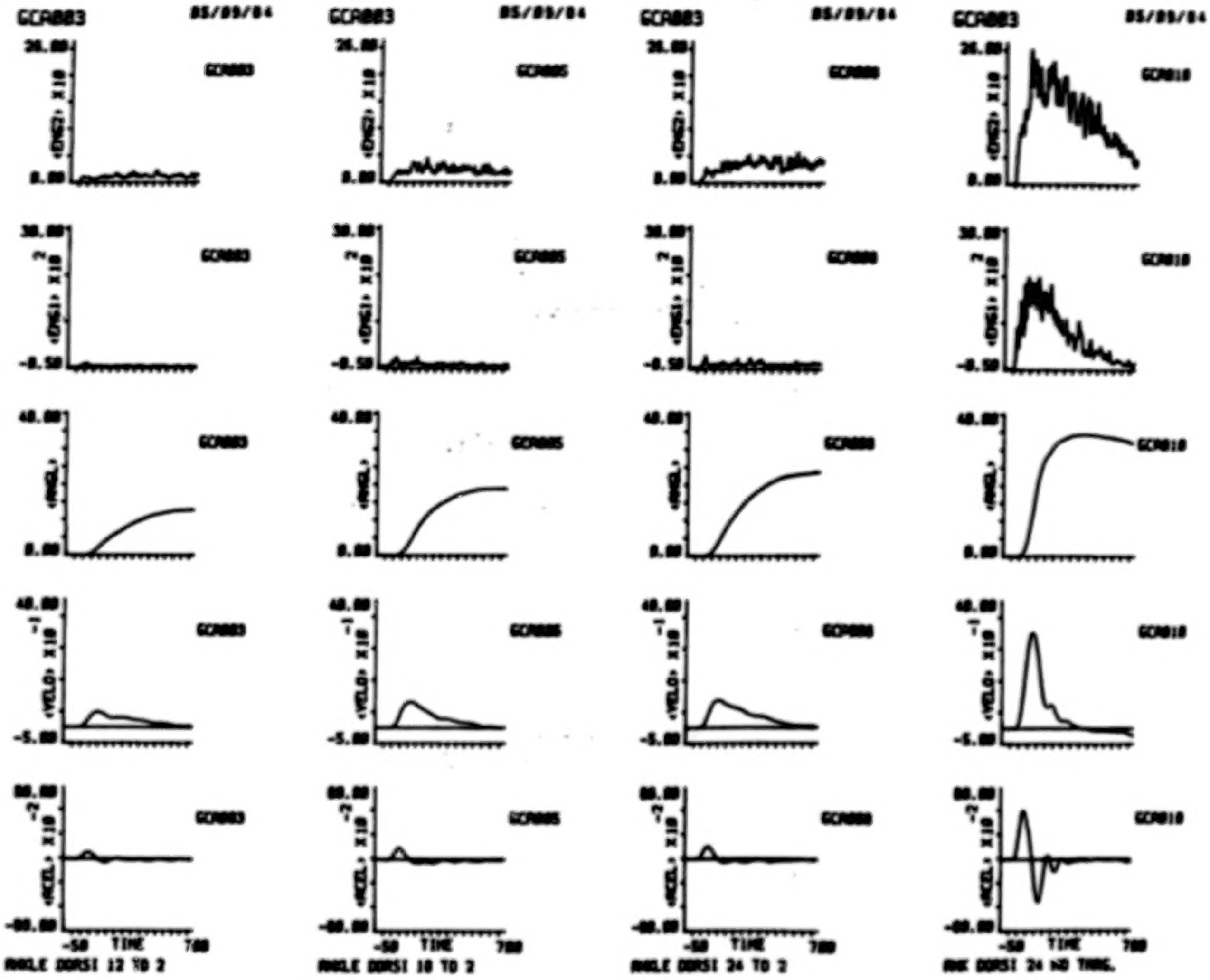


Figure 3 contains the same variables as in Figure 1. The only difference is that the data is from a different subject.

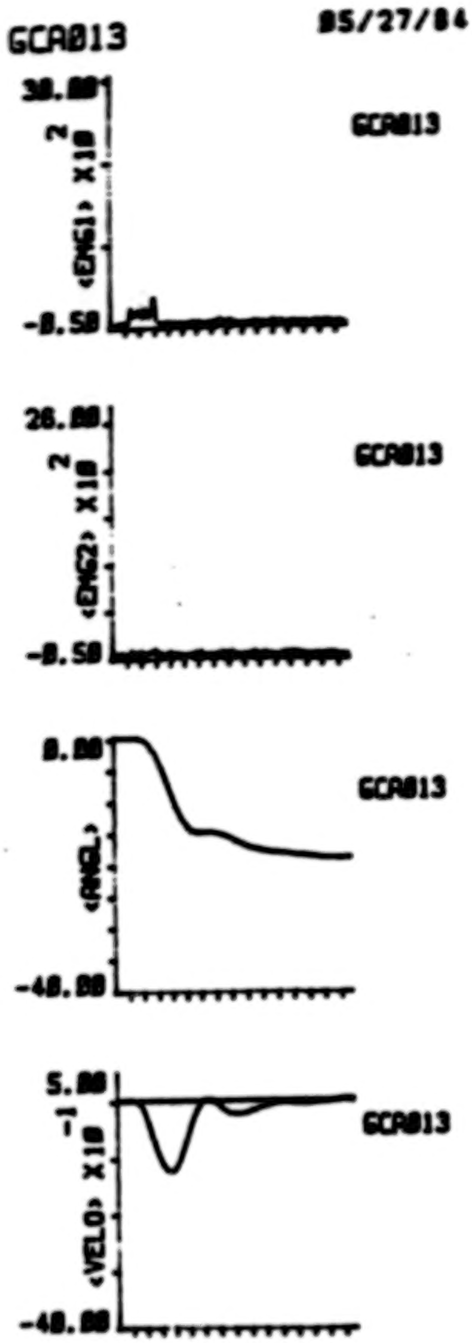


Figure 4 shows the average data from the same subject as in Figure 3. EMG1 (soleus) is plotted on the same scale in both figures. EMG2 (tibialis anterior) is 10 times larger than that in Figure 3.

target. Again, it is interesting to compare dorsi and plantar flexion movements. If we compare GCA005 (Figure 3) with GCA013, it can be seen that EMG1 (soleus) has similar activation as both antagonist (GCA005) and agonist (GAC013). However, EMG2 (tibialis anterior) has at least ten times more activation when it is acting as the agonist (GCA003).

Even when rapid movements are made, the patterns of some individuals should be considered as discrete bursts only with considerable reservation. For example, Figure 5 demonstrates multiple, successive bursts of agonist and antagonist activation which are coactivated. The data are an individual trial of a subject making a 75 degree elbow extension movement to a 3 degree target. This example highlights that extreme caution must be taken in trying to characterize two continuous patterns of EMG activity by a small set of parameters describing durations and intensities. The fact that so much reciprocal and coactivated activity is occurring implies to us that intrinsic muscle compliances make major contributions to joint torques; without understanding these torques the EMG patterns cannot be explained. This is especially true when rapid movements are made and effective load compensation is required (Grillner, 1972).

Although the EMG pattern is more complicated than two or three fixed bursts of activation, we attempted to make measurements from the first agonist and first antagonist burst. This was done to establish how the first part of the signal is scaled with respect to distance and target size. Specifically, is the duration of the first burst agonist burst independent of distance moved?

Effect of distance and target size on ankle agonist duration

The duration of the first agonist burst was determined by visual inspection and computer assisted measurements of the individual trials. These trials were then averaged and the results are presented in Table 1. The data are the combined average of ten trials for all seven subjects. A repeated measures analysis of variance was performed on the data. There was a significant effect of distance moved F (2,24) = 11.03, $p < .05$ and also for target size, F (2,24) = 3.69, $p < .05$. There was no interaction. This suggests that EMG duration increased when longer movements were made and also when movements were made to larger targets. It is a finding which causes a problem for a strict interpretation of the pulse-step model since these are opposing effects.

Effect of distance and target size on ankle agonist intensity

Intensity was determined by integrating the area under the EMG curve corresponding to duration. The rectified and filtered activity are presented in arbitrary units in Table 2. The same statistical test was adopted. Distance moved was significant, F (2,12) = 17.7, $p < .05$, but target size was not. This suggests that longer movements are made by increasing the intensity of muscle contraction. The conclusion for target size will be treated with caution since inspection of the data suggests consistent increases in intensity across target width for all three distances moved. One possible reason for the statistical non-significance of the intensity data is the inconsistency across conditions of one of our seven subjects. Table 3 shows normalized intensity (intensity divided by duration) and supports the idea

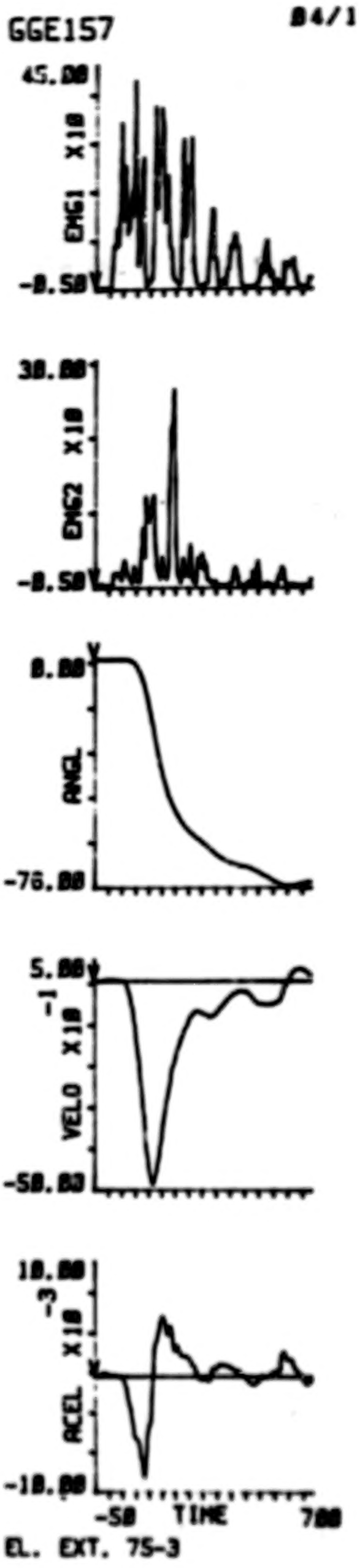


Figure 5 depicts an individual trial of a 75 degree elbow extension movement to a 3 degree target. EMG1 refers to the triceps and EMG2 to the biceps muscle. Angle is in degrees, velocity in degrees per second and acceleration in degrees per second per second.

TABLE 1

DURATION OF FIRST AGONIST EMG BURST (ms)

		Distance (deg)			
		12	18	24	Ave
Target	2	67	84.7	94	81.9
Size (deg)	4	72.6	87.0	93.4	84.3
	8	82	83.6	97.1	87.6
	Ave	73.9	85.1	94.8	

TABLE 2

INTENSITY OF FIRST AGONIST EMG BURST

		Distance (deg)			
		12	18	24	Ave
Target	2	23.4	36.8	51.6	37.3
Size (deg)	4	26.8	42.1	53.7	40.9
	8	31.5	47.4	64.5	47.8
	Ave	27.3	42.1	56.6	

that increased intensity is not caused only by an increase in the duration of the contraction but by scaling of intensity and duration. The exact form of the function remains to be elucidated. We need to manipulate a wider range of distances and target sizes.

Effects of distance and target size on ankle antagonist duration

The findings of the antagonist muscle are very similar to those of the agonist muscles except that there was no effect of target width on EMG duration. The data are presented in Table 4.

Effects of distance and target size on ankle antagonist intensity

The results are very predictable from the data already presented. Intensity increased as a function of distance moved, but not target size. See Table 5 for the intensity data and Table 5 for the normalized data.

DISCUSSION

This study suggests that it is unlikely that all movements are initiated by a pulse of constant duration. Instead, it seems that movements are initiated by an agonist burst which is scaled both in the amount of activation and the duration of activation according to either distance, target size, velocity, or a combination of factors. The number of bursts varies considerably and further research is required to establish: 1) which factors affect the pattern of the signal and 2) how different patterns produce movement trajectories.

The antagonist burst can occur at any time following the agonist burst and is also scaled. Further studies are required to establish the course of the scaling relationship between the agonist and antagonist muscles and ascertain whether they are controlled independently or as a unit.

ACKNOWLEDGMENT

This work was partially supported by the NIH Grants NS-15630 and AM-33189 and NSF Grant IESE-8212067.

REFERENCES

Agarwal, G. C. and Gottlieb, G. L., Oscillation of the human ankle joint in response to applied sinusoidal torque on the foot. Journal of Physiology (London), 228, pp. 151-176, 1977.

Agarwal, G. C. and Gottlieb, G. L., Mathematical modeling and simulation of the postural control loop: Part I. CRC Critical Reviews in Biomedical Engineering, 8, pp. 93-134, 1982.

Angel, R. W., Electromyography during voluntary movement: the two burst pattern. Electroencephalography and Clinical Neurophysiology, 36, pp. 493-498, 1974.

Brown, S. H. C. and Cooke, J. D., Amplitude and instruction-dependent modulation of movement-related electromyogram activity in humans. Journal of

TABLE 3

NORMALIZED AGONIST INTENSITY (Intensity/Time X 100)

		Distance (deg)			
		12	18	24	Ave
Target	2	34.9	43.4	54.9	44.4
Size (deg)	4	37.0	48.4	57.5	47.6
	8	38.4	56.7	66.4	53.8
	Ave	36.8	49.5	59.6	

TABLE 4

DURATION OF FIRST ANTAGONIST EMG BURST (ms)

		Distance (deg)			
		12	18	24	Ave
Target	2	44.1	58.4	59.4	54
Size (deg)	4	49.6	51.8	66.4	55.9
	8	53.5	57.2	65.1	58.6
	Ave	49.0	55.8	63.7	

TABLE 5

INTENSITY OF FIRST ANTAGONIST EMG BURST

		Distance (deg)			
		12	18	24	Ave
Target	2	2.2	3.2	3.9	3.1
Size (deg)	4	2.6	3.2	4.7	3.5
	8	3.1	3.5	5.7	4.1
	Ave	2.6	3.3	4.8	

TABLE 6

NORMALIZED ANTAGONIST INTENSITY (Intensity/Time X 100)

		Distance (deg)			
		12	18	24	Ave
Target	2	5.0	5.5	6.6	6.0
Size (deg)	4	5.2	6.2	7.1	6.2
	8	5.8	6.2	8.8	6.9
	Ave	5.3	6.0	7.5	

Physiology (London), 316, pp. 97-107, 1981.

Corcos, D. M., An analysis of the mechanisms involved in the control of rapid, single limb, uni-directional movements. Ph.D. dissertation, University of Oregon, 1982.

Descartes, R., Treatise of Man, 1637. (Translation by T. S. Hall, Harvard Univesity Press, Cambridge, MA, 1972.)
Enoka, R. M., Muscular control of a learned movement: the speed control system hypothesis. Experimental Brain Research, 51, pp. 135-145, 1983.

Fitts, P. M., The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology, 47, pp. 381-391, 1954.

Fitts, P. M. and Peterson, J. R., Information capacity of discrete motor responses. Journal of Experimental Psychology, 67, pp. 103-112, 1964.

Freund, H. J. and Budingen, H. J., The relationship between speed and amplitude of the fastest voluntary contractions of human arm muscles. Experimental Brain Research, 31, pp. 1-12, 1978.

Freund, H. J., Motor unit and muscle activity in voluntary motor control. Physiological Reviews, 63, pp. 387-436, 1983.

Ghez, C., Contribution to rapid limb movement in the cat. In H. Asanuma and V. J. Wilson (eds.), Integration in the Nervous System, Igaku Shoin, Tokyo, 1979.

Grillner, S. The role of muscle stiffness in meeting the changing postural and locomotor requirements for force development by the ankle extensors. Acta Physiologica Scandivica, 86, pp. 92-108, 1972.

Hallet, M. and Marsden, C. D., Ballistic flexion movements of the human thumb. Journal of Physiology, 294, pp. 33-50, 1979.

Hill, A. V., The abrupt change from rest to activity in muscle. Proceedings of the Royal Society, B76, pp. 136-195, 1938.

Hogan, N., Adaptive control of mechanical impedance by coactivation of antagonist muscles. IEEE Transactions on Automatic Control (in press).

Lestienne, F., Effects of intertial load and velocity on the braking process of voluntary limb movements. Experimental Brain Research, 35, pp. 407-418, 1979.

Mustard, B. E. and Lee, R. G., Evidence for central programming of the antagonist EMG burst during rapid voluntary movements. Neuroscience Abstacts, 9, p. 1033, 1983.

Schmidt, R. A., Zelaznik, H., Hawkins, B., Frank, J. S. and Quinn, J. T., Motor-output variability: a theory for the accuracy of rapid motor acts. Psychological Review, 86, pp. 415-451, 1979.

Smith, A. M., The coactivation of antagonist muscles. Canadian Journal of Physiology and Pharmacology, 59, pp. 733-747, 1981.

Smith, O. J. M., Nonlinear computations in the human controller. IEEE Transactions on Biomedical Engineering, BME-9, pp. 125-128, 1962.

Tilney, F. and Pike, F. H., Muscular coordination experimentally studied in relation to the cerebellum. Archives of Neurology and Psychiatry, 13, pp. 289-334, 1925.

Wadman, W. J., Denier van der Gong, J. J., Geuze, R. H. and Mol, C. R., Control of fast goal-directed arm movements. Journal of Human Movement Studies, 5, pp. 3-17, 1979.

Wallace, S. A., An impulse-timing theory for reciprocal control of muscular activity in rapid, discrete movements. Journal of Motor Behavior, 13 pp. 144-160, 1981.

Wallace, S. A. and Wright, L., Distance and movement time effects on the timing of agonist and antagonist muscles: a test of the impulse-timing theory. Journal of Motor Behavior, 14, pp. 341-352, 1982.

Waters, P. and Strick, P. L., Influence of "strategy" on muscle activity during ballistic movements. Brain Research, 207, pp. 189-194, 1981.

C-3

N85
14552

UNCLAS

N85 14552 D17

COLOR AND GREY SCALE IN SONAR DISPLAYS

K.-F. Kraiss, K.-H. Küttelwesch

Forschungsinstitut für Anthropotechnik
(FGAN/FAT)
Königstr. 2
D-5307 Wachtberg-Werthhoven
F.R. of Germany

- Informal[illegible] paper
- Submitted to the Regular Annual Manual

Objectives and Background:

In spite of numerous publications [1] it is still rather unclear, whether color is of any help in sonar displays. The work presented here deals with a particular type of sonar data, i.e., LOFAR-grams (low frequency analysing and recording) where acoustic sensor data are continuously written as a time-frequency plot. The question to be answered quantitatively is, whether color coding does improve target detection when compared with a grey scale code.

In order to enable an experimental evaluation, synthetic test pictures have been developed that show vertical target lines in front of noisy background on a high precision TV-screen. Experimental variables were the signal-to-noise ratio, background noise level, background intensity, and 7 variations in color or grey scale. The generation of reasonable scales turned out to be a major problem, since the steps must be equally spaced in terms of chrominance and luminance. This goal was finally achieved using the

PRECEDING PAGE BLANK NOT FILMED

photo-colorimetric space concept as proposed by Galves and Brun [2,3]. Figure 1 below shows representative dimensions of this space for color TV-monitors.

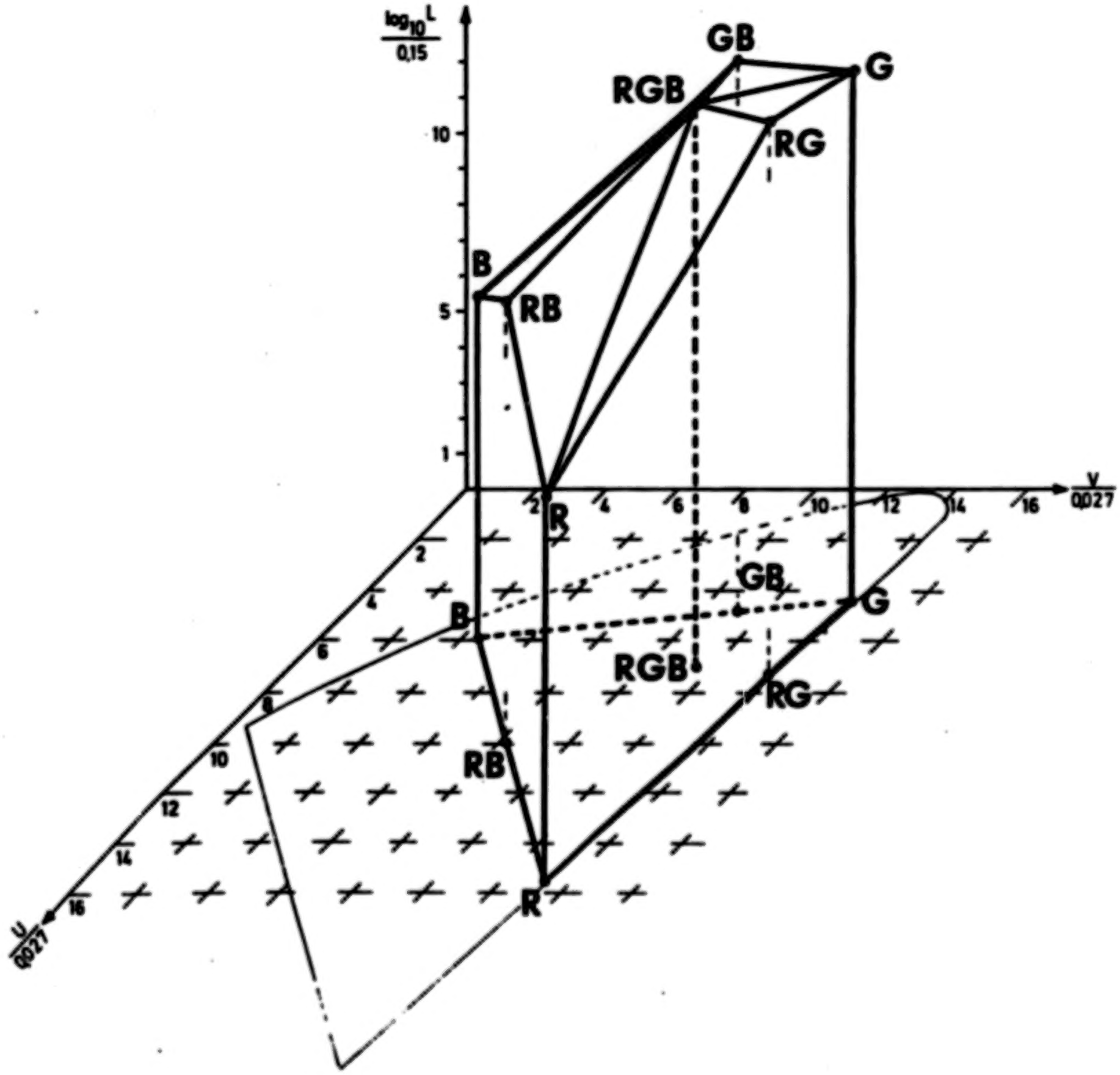

$\frac{\log_{10} L}{0{,}15}$
10
5
1
GB
G
RGB
RG
B
RB
R
$\frac{v}{0{,}027}$
2
4
6
8
10
12
14
16
$\frac{u}{0{,}027}$
2
4
6
8
10
12
14
16
GB
G
B
RGB
RG
RB
R

Fig. 1: Photo-colorimetric space

A computer algorithm has been developed that automatically generates scales within this space, e.g., from blue (20 cd/m²) over red (40 cd/m²) to white (140 cd/m²), and in addition indicates the number of jnd's available on this path. A prerequisite to run this algorithm is a nomogram (see Fig. 2) that must be established for the particular TV-monitor in use.

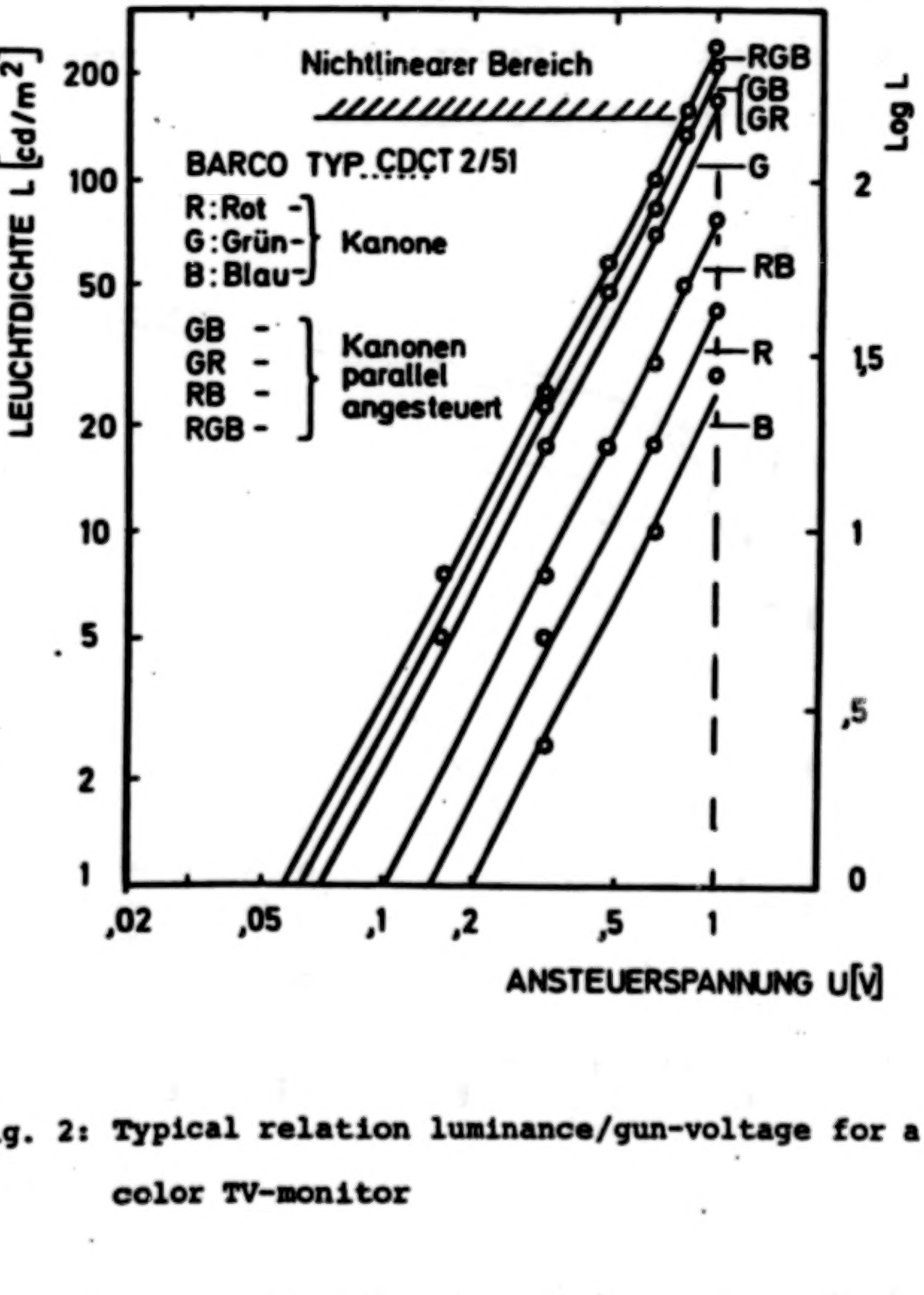


Fig. 2: Typical relation luminance/gun-voltage for a color TV-monitor

Results:

A signal detection experiment has been performed using the test pictures described above with 7 grey scale and color codes. Results are presented as receiver-operating characteristics (ROC-curves), where the background noise level turns out to be the main factor. (The specimen presented in Figure 3 is taken from a pilot study).

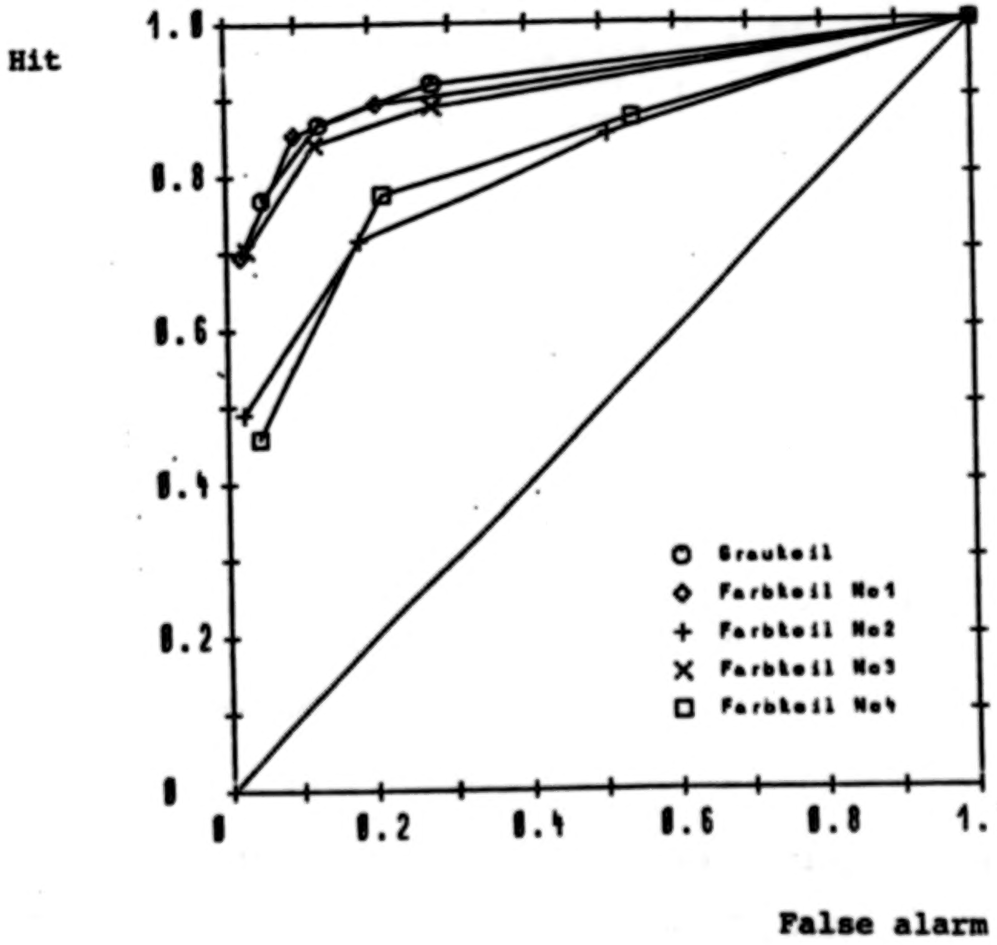


Fig. 3: ROC-curves for various codes and a particular background noise level.

The data show significant differences in ROC-performance for the selected codes. In addition it turned out, that the background noise level affects the performance dramatically for some color codes, while others remain stable or do even improve. This paper presents generally valid rules, how to generate useful color scales for this particular application. To our knowledge, comparable quantitative data on color coding have not been published elsewhere so far.

References:

[1] French, R.S.; Grey-Scale Versus Color Coding of Acoustic Data Images; NOSC, S. Diego, TR 207, 1978.

[2] Galves, J.-P., Brun, J.; Color and Brightness Requirements for Cockpit Displays: Proposal to Evaluate their Characteristics. Twenty-ninth Agard Avionics Panel Technical Meeting.

[3] Christiansen, P.; Design Considerations for Sunlight-Viewable Displays. Proc. of the SID, Vol. 24/1, 1983.

N85
14553

UNCLAS

N85 14553 D18

MANUAL-CONTROL ANALYSIS APPLIED TO THE MONEY-SUPPLY CONTROL TASK

R. C. Wingrove
NASA Ames Research Center, Moffett Field, CA 94035

ABSTRACT

The recent procedure implemented by the Federal Reserve Board to control the money supply is formulated in the form of a tracking model as used in the study of manual-control tasks. Using this model, an analysis is made to determine the effect of monetary control on the fluctuations in economic output. The results indicate that monetary control can reduce the amplitude of fluctuations at frequencies near the region of historic business cycles. However, with significant time lags in the control loop, monetary control tends to increase the amplitude of the fluctuations at the higher frequencies. The study outlines how the investigator or student can use the tools developed in the field of manual-control analysis to study the nature of economic fluctuations and to examine different strategies for stabilization.

LIST OF SYMBOLS

G	control gain
GNP	Gross National Product
j	$\sqrt{-1}$
K	constant
M	money supply based on M1 (currency and all checking accounts), dollars
m	rate of growth in money supply, 100 d(log M)/dt, %/yr
r_m	random monetary disturbance
r_x	random nonmonetary disturbance
T	time delay, yr
t	time, yr
t_r	return time, yr
X	real GNP, constant dollars
x	rate of growth in real GNP, 100 d(log X)/dt, %/yr
ζ	damping ratio (nondimensional)

PRECEDING PAGE BLANK NOT FILMED

ω	oscillatory frequency, rad/yr
ω_n	natural frequency, rad/yr
Δ	variation about long-term trend
$(\ )_o$	long-term trend
$(\ )_t$	target

I. INTRODUCTION

Adam Smith in [1] observes a natural tendency of free-market societies to balance supply and demand through exchange using the price mechanism. This self-regulating system is dynamic with fluctuations caused by a diversity of effects from individuals, business, government, and nature. Within this system, the wealth of a nation is in the output of goods and services rather than in the supply of money used for exchange. Although the money supply may be neutral with respect to real output in the long run, there can be short-term effects of money on real output. David Hume in [2] observes changes in real output as the economy self-regulates to new overall price levels that are caused by changes in the money supply. Following the theoretical framework of Milton Friedman [3], a model for these economic dynamics was developed by Dean Taylor in [4]. This model was used by the author in [5] to study how output fluctuations, termed "business cycles," have been influenced by the historical movements in the money supply. This paper uses the model further to study how the nature of output fluctuations can be influenced by the recent change in the procedure used by the Federal Reserve Board to control the money supply.

As described in [6]-[9], a new operating procedure to control the money supply was implemented by the Federal Reserve Board in October 1979. The overall policy is to target just enough growth in the money supply to finance a sustainable growth in economic output, and not so much as to feed inflation [10]. The operating procedure is implemented from the trading desk at the Federal Reserve Bank in New York to manually control the actual path in the measured money supply toward the target path in money supply. This control task, which seeks to reduce deviations between the measured and the target values, represents one form of a so-called "tracking task" in the field of manual control [11], [12]. To provide a means to study the dynamic properties of the new monetary-control procedure, this paper will utilize methods that have been used extensively in manual-control analysis.

The paper is organized in three sections. The first section introduces the model that will be used to represent the dynamic relationship between the growth of money supply and the growth of real Gross National Product (GNP). The next section considers the recently implemented procedure used to control the money supply, and formulates this control task in the form of a tracking model. The last section then combines the models from the previous sections, and studies the effects of monetary control on the nature of the fluctuations in the growth of real GNP.

II. ECONOMIC DYNAMICS

Following the theoretical framework of Friedman [3], a model for the economic dynamics was developed by Taylor [4] as a linear second-order differential equation. As shown in Fig. 1, the input variable is the growth of money supply m and the output variable is defined as $\Delta x = x - x_0$, where x is the growth of real GNP and x_0 is the long-term growth trend. The parameters used to characterize the self-regulating dynamics are the natural frequency ω_n and damping ratio ζ, and a gain K. Representative values for these parameters, based on the period 1950-1981, are $\omega_n = 1.5$ rad/yr, $\zeta = 0.8$, and $K = 2$ yr. These numerical values were determined through parameter identification in [5].

The capability of this relatively simple model to describe the monetary contributions to the growth of real GNP for the period 1950-1981 is illustrated in Fig. 1. The upper chart in Fig. 1 presents the measured growth of money supply (m) over this 32-yr period. The lower chart presents the measured growth of real GNP (x) along with the calculated output from the model ($\hat{x}$). In the lower chart, the long-term trend (x_0) is indicated by the representative value of 3%/yr.

As shown in the lower chart of Fig. 1, the growth of real GNP tends to fluctuate around the long-term trend. The contributions to the fluctuations include the monetary effects as computed from the model along with nonmonetary effects. Nonmonetary effects during this time period include the 1964/65 tax cut, the 1968/69 tax surcharge, and the 1973/74 oil shock. Nonmonetary effects of this type and their interactions with the economic dynamics are discussed further in [5].

Background literature [13]-[21] points out that, historically, the growth in money supply has moved in conjunction with business activity. The study in [5] analyzes how these procyclical movements in the money supply reduce the damping of the overall (closed-loop) system dynamics. Such lightly damped dynamics when subjected to random disturbances tend to produce business cycles.

Figure 2 presents a histogram of past business cycles (1857-1980) in comparison with the representative value for the natural frequency, $\omega_n = 1.5$ rad/yr. We observe that most business cycles are gathered about the natural frequency in a region from about 0.5 to 3 rad/yr. This region of cyclical movements will be termed the business-cycle frequency range in later discussions.

III. MONETARY CONTROL

In October 1979, a new operating procedure was implemented to control the money supply. As described in [6]-[9], the concept uses return paths to reduce any error between the target trend in money supply and the measured trend in money supply. The return paths, sketched in the upper part of Fig. 3, depend upon a chosen value for what is termed the "return time," t_r [9]. In the context of manual-systems analysis, the inverse of the return time $(1/t_r)$ is equivalent to the value for the "gain" used to reduce the error. A mathematical model for this monetary control task is presented in the lower part of Fig. 3. This formulation is based on the standard crossover model that has been used extensively in the field of manual-control analysis [11]. In this formulation, the monetary-control gain $(1/t_r)$ is represented by G. The time required to ascertain the error in the money supply and to implement the policy[1] is represented by a time delay T. Disturbances in the growth of money supply are represented by the residual r_m.

Representative values for the parameters G and T have been estimated from some of the past data that are shown in Fig. 4. The upper graph presents the target growth in money supply given in terms of year-to-year target ranges for M1. The lower graph presents the measured growth of money supply based on the month-to-month changes (annualized) in M1. Representative tracking data are illustrated after implementation of the new operating procedure. These data start in the spring of 1980 (following a fairly strong, downward monetary disturbance apparently caused by credit restrictions at that time) and continue through mid-1982 (at which time the control of M1 was relaxed because of financial innovations and regulatory changes upcoming in later months). Using these representative time-series data, the parameters G and T were estimated by a two-parameter search method as in [12]. Values for the gain G were estimated to be approximately 3/yr to 4/yr, which appear to be consistent with the values as described in [9]. The values for the time delay T were estimated to be about 0.1 yr to 0.2 yr. These values appear reasonable based on a description of the operating procedure as discussed in [6]-[9].

The general trend of the targets during the period from 1979 through mid-1982 was to reduce the growth in money supply by about 1/2% for each year. A trend line in the measured growth of the money supply during this period is shown in the lower graph in Fig. 4. We observe that the trend matches quite well the trend in the targets from 1979 through mid-1982. We also note large-amplitude, shorter-period fluctuations about the trend in the measured growth of money supply. These frequency-response characteristics of the monetary-control task will be examined further in the next section.

---

[1] Implementation is primarily through the buying or selling of U.S. securities on the open market [6]-[9]. The time delay T includes the time required for open-market operations to affect the growth of money supply.

IV. ANALYSIS

Using the monetary-control model, this section first examines the effects of monetary control on the nature of the fluctuations in the growth of money supply. The monetary-control model is next combined with the economic model from section II to study the effects of monetary control on the nature of the fluctuations in the growth of real GNP. An example of policy feedback is then considered, leading to a final note on the extensions of this analysis approach.

A. Monetary-control dynamics

The time-response characteristics of the monetary-control model are illustrated by the graph in Fig. 5. This graph shows the time history in the growth of money supply m responding to an impulse in the disturbance r_m. The gain G is fixed at 3/yr, representing a return time of $t_r = 0.333$ yr (4 mo). The curves in this graph have been computed for different values of the time delay T. For the computed curve with $T = 0$, the response is an exponential decay with a time constant equivalent to the return time t_r. The other curves show that the effect of the time delay T is to increase the tendency to overshoot and produce an oscillatory response. Overshoot occurs (for $G = 3$/yr) when the time delay T is greater than about 0.13 yr (1.5 mo). We next examine the response for the more general range of possible disturbances r_m.

The frequency response characteristics of this tracking model are illustrated by the graph in Fig. 6. This graph shows the multiplier $|m/r_m|$ representing the fluctuations in the growth of money supply (m) caused by random monetary disturbances (r_m). The curves in this graph have been computed for different-value time delays T, with the gain G fixed at 3/yr. The curves show that the amplitude of the fluctuations in m, owing to r_m, are reduced primarily at those frequencies below the numerical value of G (the gain G is equivalent to the bandwidth, or crossover frequency). The effect of the time delay T is to increase the amplitude of the fluctuations in m at the higher frequencies (for reference: the money-supply tracking data in Fig. 4 appear to show an increased amplitude of the fluctuations in m at the higher frequency range of approximately 6 to 10 rad/yr; i.e., cycle periods on the order of a year or less).

B. Monetary control and economic dynamics

Using this tracking model, we can examine the effects of monetary control on the general nature of economic activity. The mechanism is shown in the flow diagram at the top of Fig. 7. This flow diagram shows the monetary-control model combined with the model from section II representing economic dynamics. The lower part of Fig. 7 presents a frequency-response graph that illustrates the effect of the monetary control on economic fluctuations. This graph shows the multiplier $|\Delta x/r_m|$ representing the amplitude of the fluctuations in the growth of real GNP (Δx) caused by monetary disturbances (r_m).

The curves in Fig. 7 have been computed for different values of the time delay T. The results indicate that with monetary control (G = 3/yr) the amplitude of the economic fluctuations are reduced primarily in the region of frequencies near the historic business cycles (near the natural frequency of the basic economic system). With the larger value of time delay T, on the order of 0.2 yr, monetary control tends to increase the amplitude of the fluctuations at the higher frequencies.

If the time delay T could be kept to a small value, these results suggest that monetary control has the potential to reduce the amplitude of the fluctuations caused by monetary disturbances. Even with monetary control, however, there still remains those economic fluctuations caused by nonmonetary disturbances (nonmonetary effects were previously discussed with Fig. 1 and are calculated as $r_x = x - \hat{x}$). One way considered to reduce the amplitude of the fluctuations caused by nonmonetary disturbances is through the use of policy feedback.

C. Monetary control and policy feedback

A formulation of the closed-loop system with a policy feedback path is presented at the top of Fig. 8. The general concept here is that the target growth in money supply (m_t) is to move countercyclical with the state of the economy (Δx). The monetary-control model is from section III, and the economic model is from section II. External disturbances include both monetary effects (r_m) and nonmonetary effects (r_x). The policy feedback is represented by a gain G' and a time delay T'.

The effects of policy feedback on the nature of economic fluctuations are illustrated by the frequency-response graphs in Fig. 8. The upper graph shows the multiplier $|\Delta x/r_m|$ representing the amplitude of the fluctuations in the growth of real GNP Δx caused by monetary disturbances r_m. The lower graph shows the multiplier $|\Delta x/r_x|$ representing the amplitude of the fluctuations in the growth of real GNP Δx caused by nonmonetary disturbances r_x.

The frequency-response curves in Fig. 8 have been computed using different values for the policy-feedback time delay T' (with monetary control fixed with G = 3/yr and T = 0.1 yr). Essentially, the results indicate that this feedback policy can reduce the amplitude of the fluctuations caused by both monetary and nonmonetary disturbances at the lower frequencies. However, this feedback policy tends to increase the amplitude of the fluctuations at the higher frequencies. The effect of the time delay T' is to increase the amplitude of the fluctuations in the upper portion of the business-cycle frequency range. The results in Fig. 8 were calculated using a representative policy-feedback gain G' = 0.4 (nondimensional). For higher values of gain G' (not shown) the effects of the time delay T' are more pronounced in amplifying the economic fluctuations in the upper portion of the business-cycle frequency range.

The results from this study are probably intuitive to those who have studied control systems or to those who have studied economic dynamics. Basically, it is difficult to achieve good stabilization if there are significant time delays in any of the control loops. Because the economy is influenced by a large number of factors, the observations are "noisy" and some amount of time is required to ascertain economic trends and proper responses. The results in this paper show that the major effect of the time delays in the control loops is to increase the amplitude of economic fluctuations at the higher frequencies.

This line of investigation can be continued by considering other values for the gains and time lags. Also, one can add dynamic elements (e.g., filtering/prediction) and additional feedback variables such as the growth in prices, interest rates, the unemployment rate as in [5], or the growth in nominal GNP as in [21]. These types of extensions appear to be promising as a means to build upon and compare different concepts about stabilization policies, and to study relationships among a broader class of dynamic variables.

V. SUMMARY REMARKS

Some results based on the application of manual-control analysis methods to the study of economic dynamics and stabilization have been presented. Economic policies and dynamics were formulated to include internal elements that respond to external random disturbances. The disturbances include both monetary and nonmonetary effects.

The procedure used to control the money supply was modeled by two parameters: a control gain and a time lag. Using this model, the study indicates that monetary control can reduce the amplitude of the fluctuations in output caused by monetary disturbances at the lower frequencies near the region of historic business cycles. With significant values of time lag, monetary control tends to increase the amplitude of the fluctuations at higher frequencies.

This study also considered a feedback policy wherein the target growth in money supply is to move countercyclical with the growth in real GNP. This countercyclical policy has the potential to reduce the amplitude of the fluctuations caused by both monetary and nonmonetary disturbances at lower frequencies. Countercyclical policy tends to increase the amplitude of the fluctuations at frequencies in the upper portion of the business-cycle frequency range.

This report outlines a fairly simple approach to mathematically represent the dynamics of the overall economy under closed-loop control. These dynamics are constructed by combining a linear second-order model (section II) with the monetary-control model (section III) and policy feedbacks (section IV). Using this framework, an investigator or student with basic skills in linear analysis can formulate and study how different stabilization strategies tend to change the dynamics of the economic system and modify the amplitude of the fluctuations that are caused by random disturbances. Future applications of this framework might be in the development of simple dynamic models to be used in prediction and forecasting.

REFERENCES

[1] A. Smith, An Inquiry into the Nature and Causes of the Wealth of Nations. Indianapolis: Liberty Classics, 1982 (originally published in 1776).

[2] D. Hume, Writings on Economics. Madison: University of Wisconsin Press, 1955 (originally in Political Discourses, published in 1752).

[3] M. Friedman, A Theoretical Framework for Monetary Analysis. New York: National Bureau of Economic Research, 1971.

[4] D. Taylor, "Friedman's Dynamic Models: Empirical Tests," J. Monetary Economics, vol. 2, pp. 531-538, 1976.

[5] R. C. Wingrove, "Classical Linear-Control Analysis Applied to Business-Cycle Dynamics and Stability," NASA TM 84366, July 1983.

[6] Federal Reserve Staff Study, New Monetary Control Procedures, vols. 1 and 2. Washington, D.C.: Board of Governors of the Federal Reserve System, Feb. 1981.

[7] Federal Reserve Bank of New York, "Monetary Policy and Open Market Operations in 1980," Federal Reserve Bank of New York, Quarterly Review, vol. 6, pp. 56-75, Summer 1981.

[8] S. H. Axilrod, "Monetary Policy, Money Supply, and the Federal Reserve's Operating Procedures." Federal Reserve Bulletin, vol. 68, pp. 13-24, Jan. 1982.

[9] P. A. Tinsley, P. von zur Muehlen, and G. Fries, "The Short-Run Volatility of Money Stock Targeting. J. Monetary Economics, vol. 10, pp. 215-237, 1982.

[10] Federal Reserve Bulletin, "Statements to Congress," vol. 70, p. 100, Feb. 1984.

[11] D. T. McRuer and H. R. Jex, "A Review of Quasi-Linear Pilot Models," IEEE Trans. Human Factors Elec., vol. HFE-8, no. 3, pp. 231-249, 1967.

[12] R. C. Wingrove, F. G. Edwards, and A. L. Lopez, "Some Examples of Pilot/Vehicle Dynamics Identified from Flight Test Records." (Presented at the 5th Annual Conference on Manual Control, Cambridge, Mass., Spring 1969) IEEE Trans. Man. Syst., vol. MMS-10, no. 4, pp. 131-132, Dec. 1969.

[13] M. Friedman and A. J. Schwartz, "Money and Business Cycles." Rev. Econ. Stat., vol. 45, no. 1, part 2, Supplement, Feb. 1963.

[14] R. C. Davis, "The Role of the Money Supply in Business Cycles." Monthly Review, Federal Reserve Bank of New York, vol. 50, pp. 63-73, April 1968.

[15] T. Mayer, Monetary Policy in the United States. New York: Random House, 1968.

[16] J. Tobin, "Money and Income: Post Hoc Ergo Propter Hoc?" Quarterly J. Econ., vol. 84, pp. 301-317, May 1970.

[17] P. A. Samuelson, "Monetarism Objectively Evaluated." Readings in Economics. New York: McGraw Hill, pp. 120-129, 1973.

[18] W. Poole, "Monetary Policies in the United States, 1965-1974. Proc. Acad. Political Sci., vol. 31, no. 4, pp. 91-104, 1975.

[19] F. Modigliani, "The Monetarist Controversy or Should We Forsake Stabilization Policies?" Amer. Econ. Rev., vol. 67, pp. 1-19, March 1977.

[20] K. Brunner and A. H. Meltzer, "Strategies and Tactics for Monetary Control," Carnegie-Rochester Conf. Series, Public Policy, vol. 18, pp. 59-104, 1983.

[21] R. J. Gordon, "Using Monetary Control to Dampen the Business Cycle: A New Set of First Principles," Nat. Bureau Econ. Res., Working Paper no. 1210, Oct. 1983.

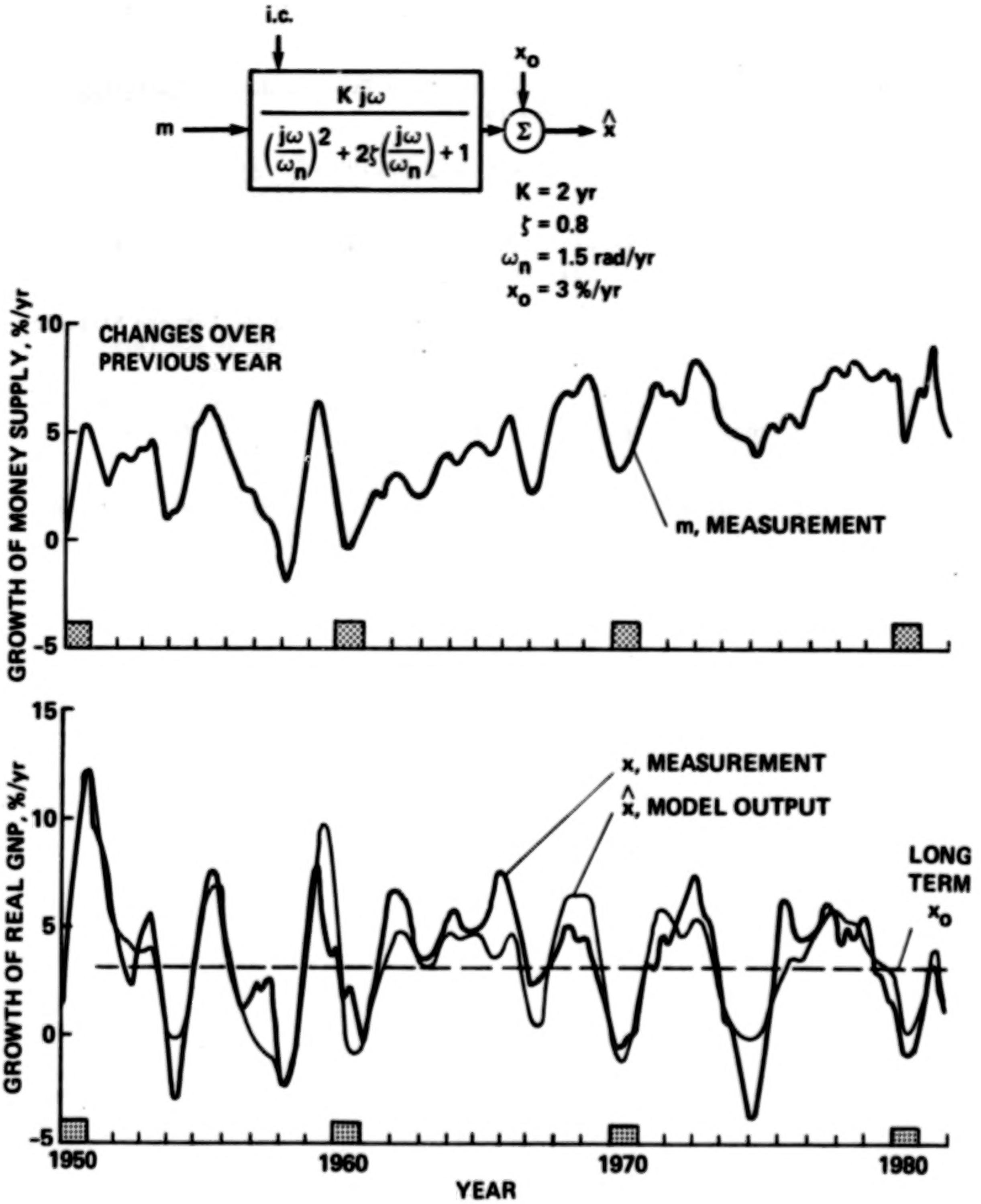


Fig. 1. Time history in the growth of real GNP x compared with the model output $\hat{x}$ and the long-term trend x_o [5].

ORIGINAL PAGE IS
OF POOR QUALITY

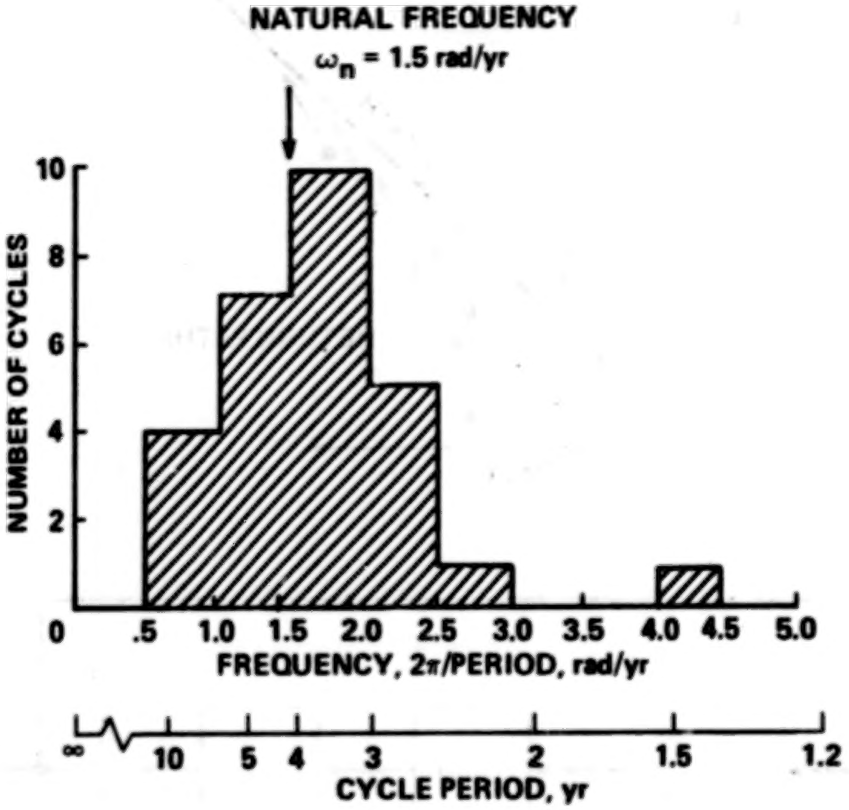


Fig. 2. Histogram of 28 business cycles, peak to peak, from 1857 to 1980 (data source: National Bureau of Economic Research, Inc.).

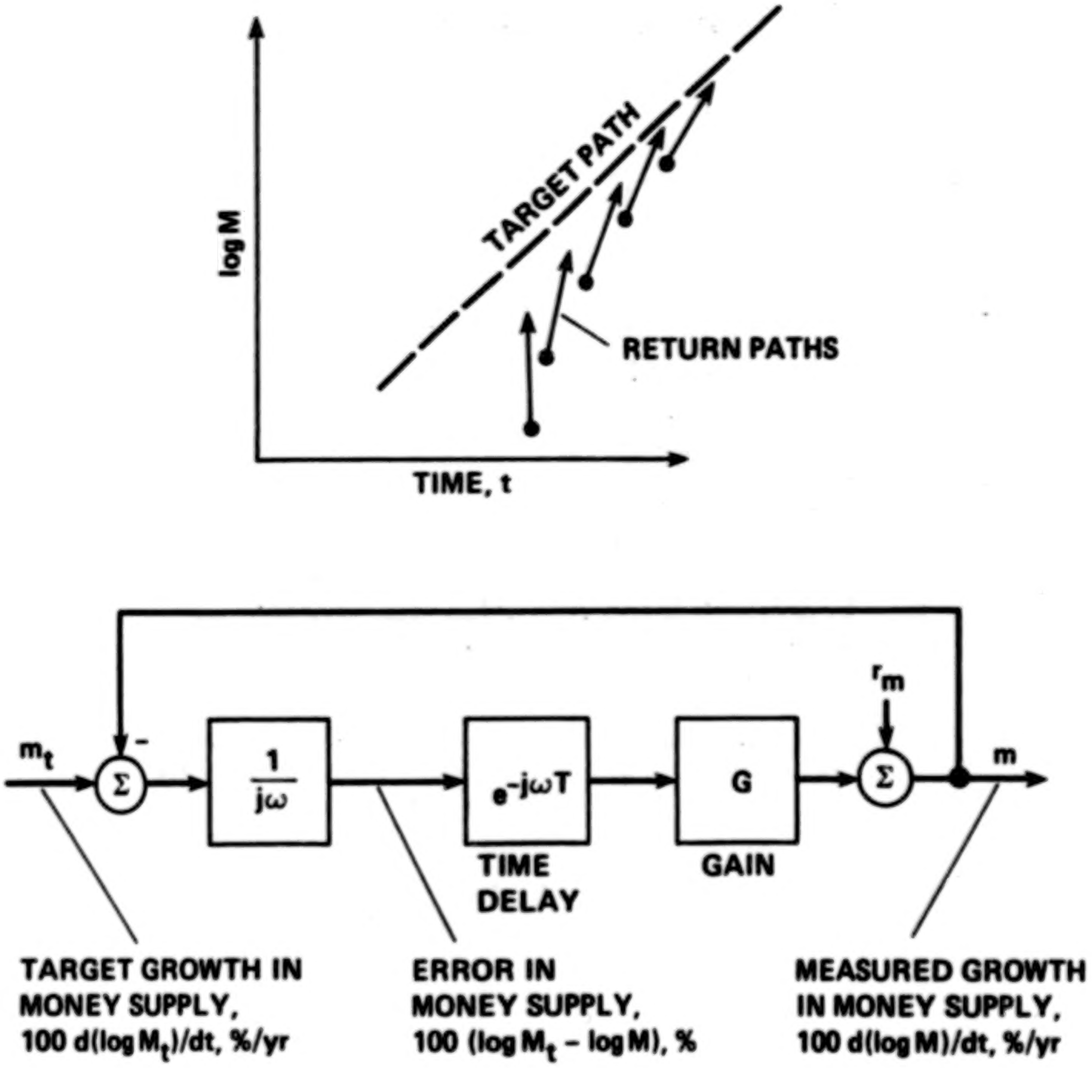

log M
TARGET PATH
RETURN PATHS
TIME, t
m_t
Σ
-
$\frac{1}{j\omega}$
$e^{-j\omega T}$
TIME DELAY
G
GAIN
r_m
Σ
m
TARGET GROWTH IN MONEY SUPPLY, 100 d(log M_t)/dt, %/yr
ERROR IN MONEY SUPPLY, 100 (log M_t – log M), %
MEASURED GROWTH IN MONEY SUPPLY, 100 d(log M)/dt, %/yr

Fig. 3. A model for the money-supply control task.

ORIGINAL PAGE IS
OF POOR QUALITY

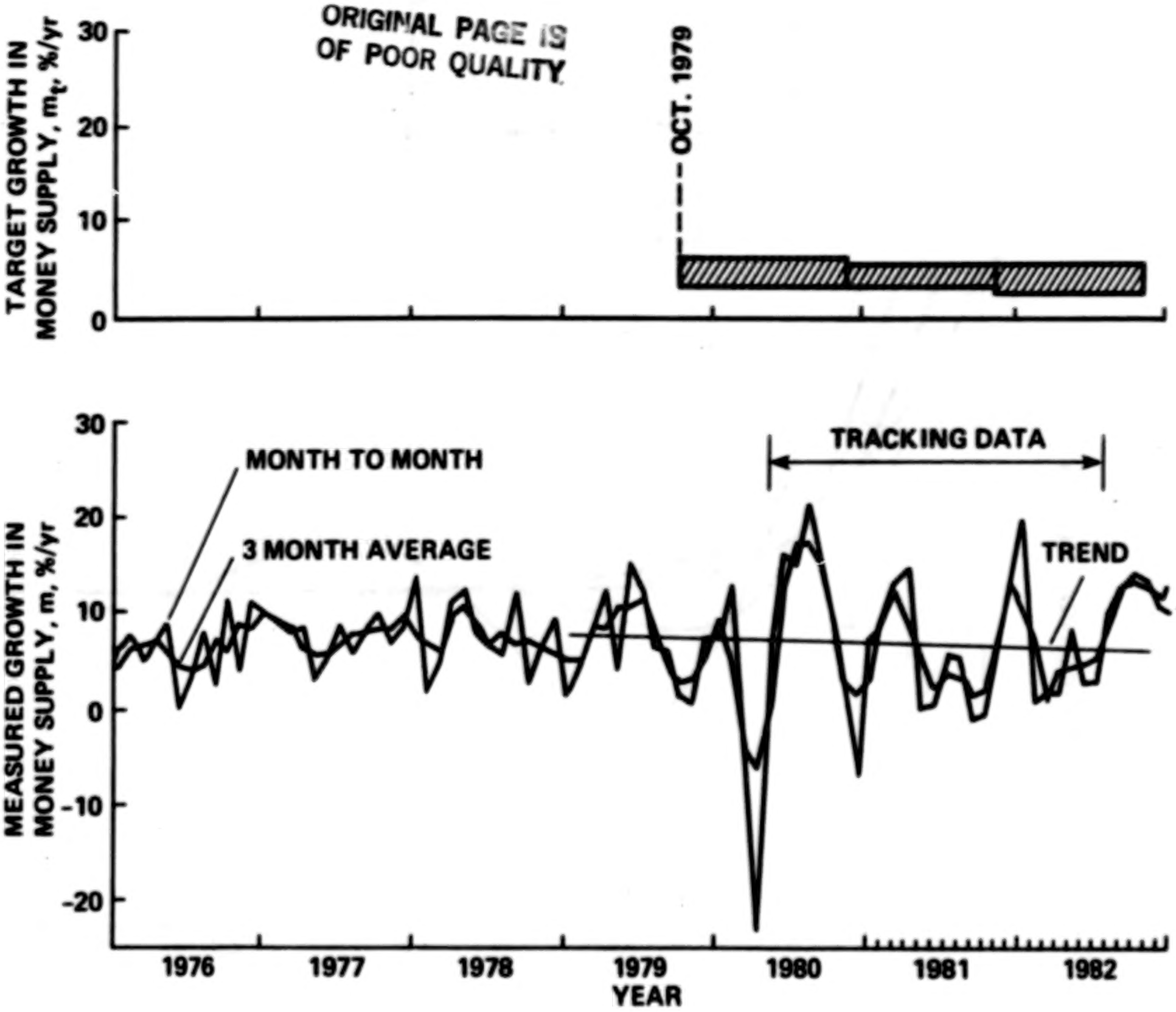


Fig. 4. Targets and measured growth rates for M1 (data source: Federal Reserve Board).

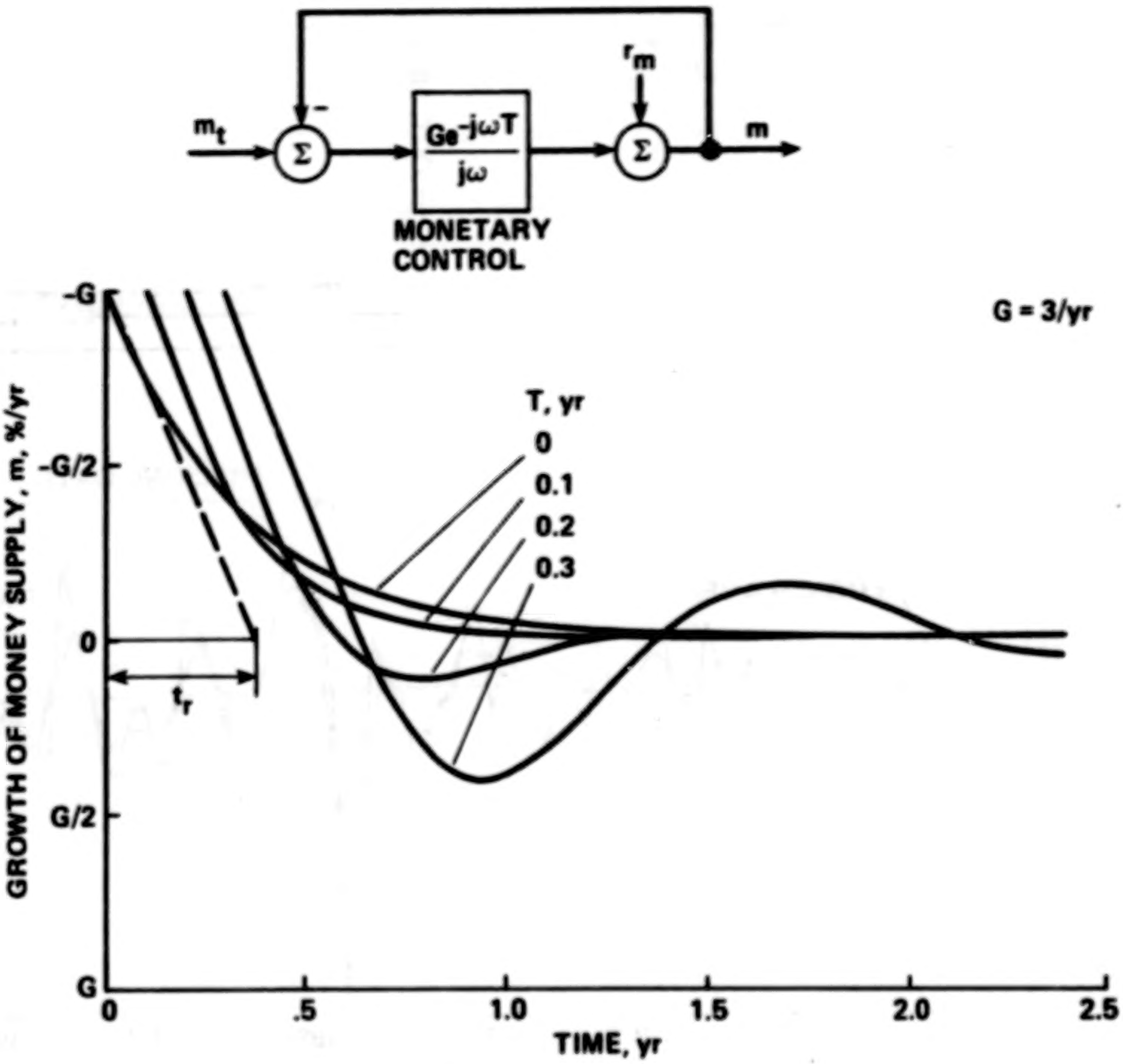


Fig. 5. Response in the growth of money supply m from an idealized impulse in the disturbance r_m.

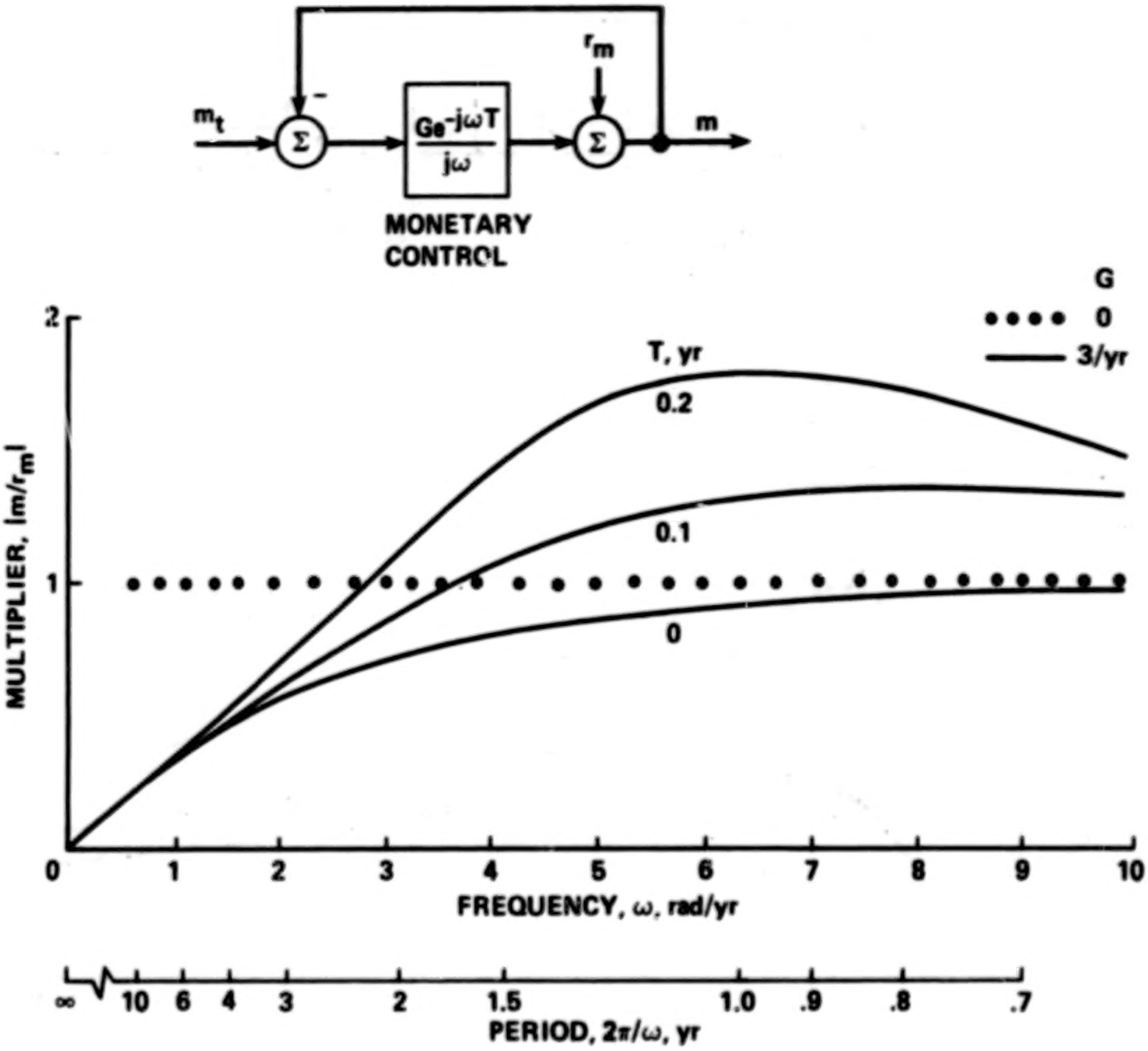


Fig. 6. Effects of monetary control on the amplitude of the fluctuations in the growth of money supply (m) caused by random monetary inputs (r_m).

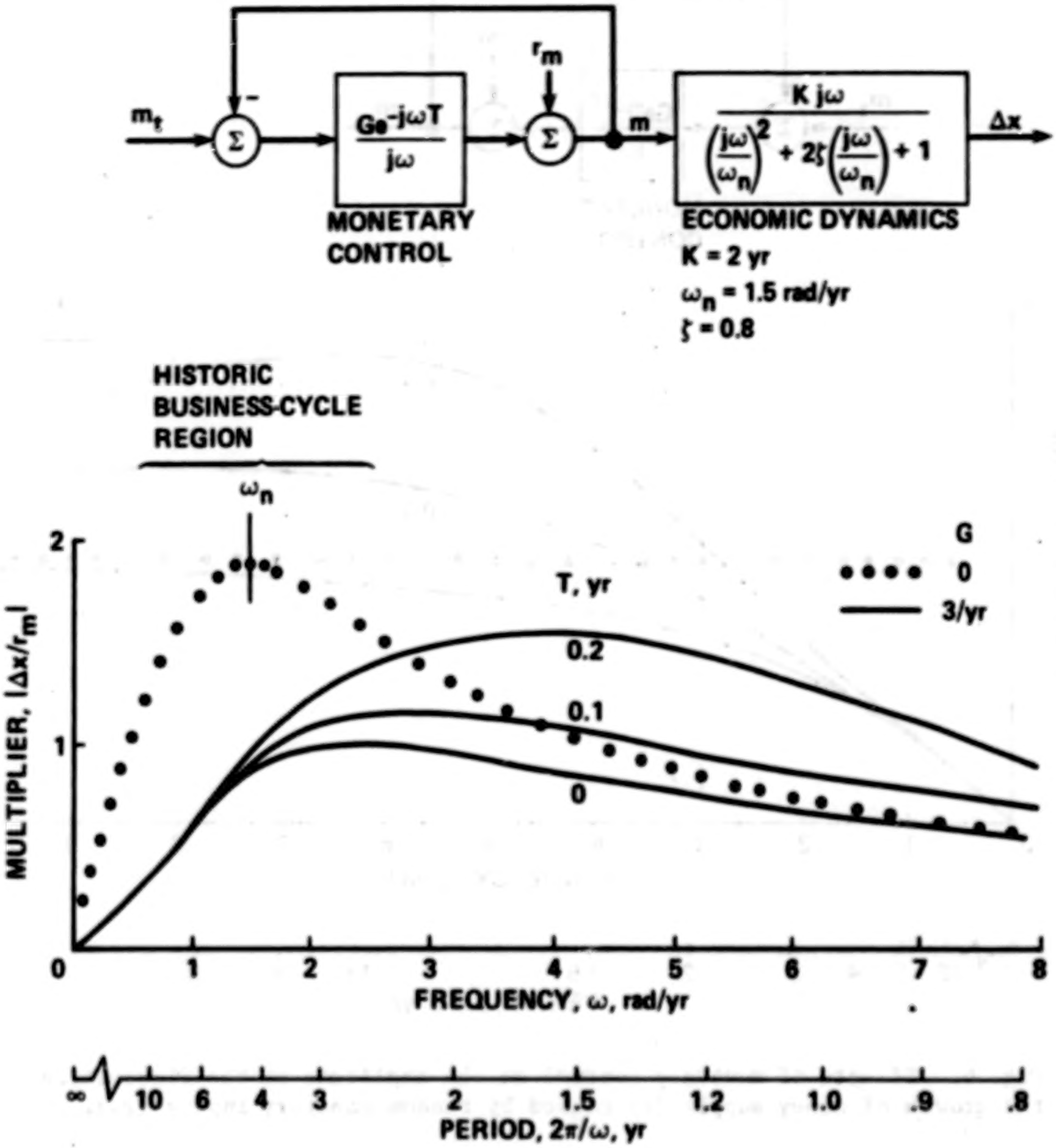


Fig. 7. Effects of monetary control on the amplitude of the fluctuations in the growth of real GNP (Δx) caused by random monetary inputs (r_m).

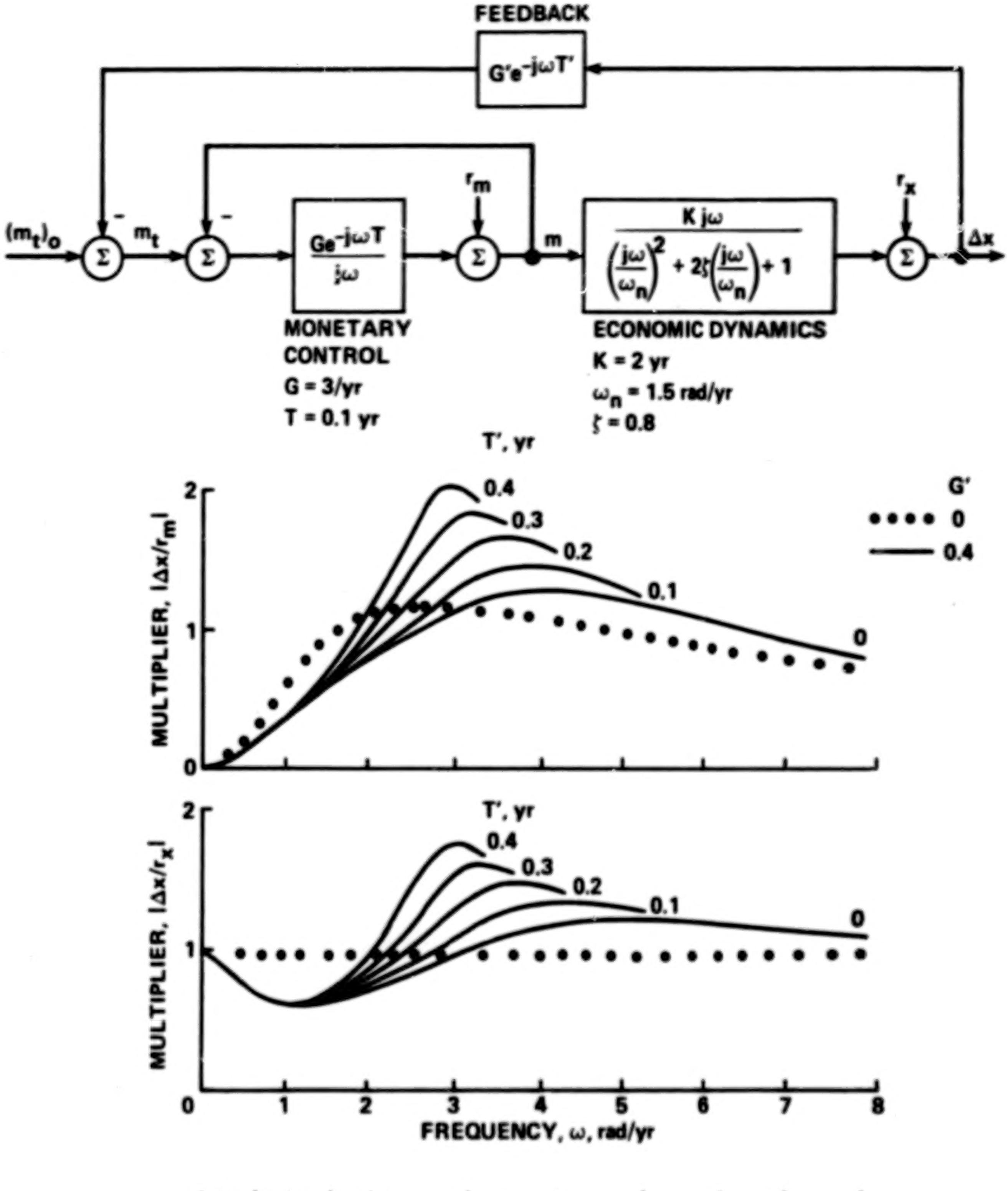


Fig. 8. Effects of policy-feedback on the amplitude of the fluctuations in the growth of real GNP (Δx) caused by random monetary inputs (r_m) and random nonmonetary inputs (r_x).

Crew Factors

N85
14554

UNCLAS

D19

N85 14554

WHAT PILOTS LIKE (AND DON'T LIKE) ABOUT THE NEW COCKPIT TECHNOLOGY

Renwick E. Curry**
Aerospace Human Factors Research Division
NASA Ames Research Center
Moffett Field, CA 94035

SUMMARY

New cockpit technology is continually required for the airlines to remain competitive, and the manufacturers respond to similar competition. A historical view of the introduction of new technology suggests that the changes have not always gone as planned, and that there have been reactions to the new technology that were not anticipated. This joint Airline/NASA study was established during the introduction of a new technology aircraft, the B-767, and had several purposes: to identify any adverse reactions to the new technology should any develop (none were found); to provide a "clearing house" of information for the airlines and pilots on the experiences during the introductory period; to provide feedback on airline training programs for the new aircraft; and to provide field data to NASA and other researchers to help them develop principles of human interaction with automated systems.

Three airlines and their pilots agreed to participate in the study. Data were obtained through more than 100 questionnaires returned by pilots, direct observation and interviews with pilots and check airmen, and attendance by a NASA observer at the ground schools of the participating airlines.

This paper reports on the pilots' perceptions of the new cockpit technology. Although the data reported in this paper were taken from the introductory experience of the B-767, it is felt that similar, if not identical, results would be obtained with any other new cockpit technology aircraft, i.e., the A310. The following conclusions have been drawn from the information collected thus far: A large majority of the pilots enjoy flying the B-767 more than the older airplanes. The pilots accept the new cockpit technology, and they choose to use it because they find it useful. The pilots are aware of the possible loss of flying skill with the presence of automation, and they hand fly (usually with flight director) to prevent this loss. There is no evidence of loss of skills from the data collected in this study. The primary points of confusion or suprise were autothrottle/autopilot interactions; the autopilot turning the "wrong way" or not capturing the course; and achieving desired results with the Flight Management System/Control Display Unit (FMS/CDU). The pilots felt training for the FMS/CDU could be improved, and they especially wanted more "hands on" experience.

**Now with Search Technology, Inc., Palo Alto, CA

INTRODUCTION

Background

New aircraft technology is continually required for the airlines and manufacturers to remain competitive. Most of the time the new technology takes the form of small, "add-on" systems to existing aircraft (such as Automated Communications and Reporting System, ACARS, or Ground Proximity Warning Systems, GPWS). Infrequently, there is a dramatic change in cockpit technology, as with the introduction of the B-767 and Airbus A310.

The Operators' View

Based on previous experience with new technology, it was expected that there would be concomitant changes required in the role of the crew, piloting techniques, procedures, and training. It was generally perceived that previous conversions to new technology did not always go smoothly; that many airlines experienced higher than expected training costs; and that some pilots experienced difficulty in the transition to the newer wide-body jets (the L-1011 and DC-10). There have been several explanations offered for this: certainly, the flight guidance systems on these aircraft are more complex than their predecessors, but it has also been noted that the captains transitioning to these aircraft had not been to school in periods of 10 to 15 years, and this may have contributed to some of the difficulties.

The Human Factors View

In many respects the technology of human factors has not kept pace with the technology of the cockpit. There is a significant body of knowledge on how to design displays and controls – material on which manual systems are based – but there is precious little material to help the human factors practitioner with the design of interfaces to complex devices. It has been felt by many observers that the performance of such systems will be determined less by traditional manual piloting skills, but more by the pilot's decision making behavior (what mode should I use?); his knowledge of the systems (is this thing working correctly?); his monitoring behavior (key strokes entered now may influence the system 5 hours later); and crew coordination (set up and monitoring of the systems and other members of the crew).

The job of the systems designer and operator is made even more complicated since many outcomes of the design and operation (such as the loss of manual skills) do not emerge until a considerable amount of experience has been gained with the new equipment. This is precisely the type of information that cannot be obtained in simulation, the traditional design tool.

In short, new human factors techniques are required to assist in the design of new cockpit technology.

Study Objectives

The objectives of the joint Airline/NASA study were the following: to identify any unanticipated side effects of the new technology that were related to safety; to provide feedback to the carriers on their training related to the new cockpit technology; to help the exchange of operational experience among carriers; to provide quantitative data on the human factors aspects of the new technology; to provide field study information for later development of human factors "principles" of automation.

Those interested in more details and other aspects of the study should consult the Technical Report (Curry, 1984).

DESCRIPTION OF THE STUDY

The study was conducted with the considerable help and cooperation of literally hundreds of individuals within the three participating airlines. The major sources of information used in the study are outlined in this section.

Ground School

The NASA observer attended the full (2 week) ground school of one airline, and 1 week periods in ground school of the other two airlines; these periods coincided with instruction of flight guidance, instrumentation, and the Flight Management System. The observer did not take the oral exam or any simulator training, but he did observe three four-hour simulator training sessions.

Pilot Volunteers

Pilot volunteers from the three participating airlines were solicited from those who attended 767 transition training. A procedure was established with the carriers whereby the anonymity of each pilot would be preserved by having him adopt an identity code number. This was necessary to establish identification for a possible second round of questionnaires. Invitations to participate in the study (a five page question and answer booklet) were prepared for each airline. Initially the invitations to participate were distributed when the pilots enrolled in the ground school for transition training. Later this was changed so that the pilots received material after their simulator training, either before or just after their initial operating experience.

Questionnaire

The primary data collection device was the questionnaire. Over 100 returns were received and 102 were used for most of the analyses. The questionnaire consisted of three parts:

Frequency of Use Table

This part was designed to to determine what features were being used by the pilots, and how frequently they used these features.

Open-Ended Questions

These questions were designed to obtain information about the features and systems that the pilots like and find useful; characteristics that they don't like; the aspects of the cockpit they would change if they could; and their opinion about the training they received.

Attitude Survey

This portion consists of 36 statements (Table 1) about the pilots' opinions on automation and flying in general, and the airplane in particular; the pilots responded on a five point "agree--disagree" Likert Scale.

Interviews and Meetings

Informal interviews were held with approximately 20 pilots and eight check pilots. Each interview lasted from one-half to one-and-one-half hours.

Progress report meetings were held at each of the three participating airlines. Attendees of these meetings consisted of representatives from fight operations management, training, line pilots, and check airmen. These progress reports seemed to have a catalytic effect, since they usually evolved into a spirited discussions among all attendees.

Cockpit Observation

The NASA observer flew as cockpit observer on one training flight (two pilots received training on this flight), two segments where a captain was receiving line training, and approximately 40 segments with line pilots operating the aircraft in normal line operation.

Internal Documentation

The airlines made available any pilot reports of irregularities or incidents that occurred.

RESULTS

Questionnaires

Respondents

A total of 104 questionnaires had been received between February 22, 1982 and July 31, 1983 (the cutoff date for the analysis). Two of the questionnaires could not be identified with a specific airline, so they have not been included in the analysis. The distribution of responses by airline, position (captain/first officer), total flying time, and time in the 767 is shown in Table 2. Perhaps the most interesting fact is that a majority of the respondents were captains, whereas our past experience has been that first officers are usually more interested in studies of this type.

Open-ended Questions

Without a doubt, the answers to the open-ended questions were the most difficult to extract and summarize, but they yielded extremely useful information. Included in this category of responses were any notations from the comment column of the frequency of use table, or comments from the pilot opinion portion of the questionnaire. These additional comments were solicited, and were quite useful.

After carefully examining 30 or so questionnaires, several categories of response began to emerge. The responses to the open-ended questions are shown in Table 3, and have been grouped into Features Liked, Features Missing or Not Liked, Points of Confusion or Surprise, and Training. Not included in these responses are those comments relating to human engineering and cockpit environmental issues, or comments regarding the implementation of a particular feature if they were not pertinent to the present study. See Curry (1984) for examples of responses to the open-ended questions.

Pilot Opinion Questionnaire

The pilots responded to 36 statements and were asked to circle one of five answers to describe how they felt about the statement: strongly agree; slightly agree; neither agree nor disagree; slightly disagree; or strongly disagree. Their responses were examined to determine if there was any correlation with the following variables: airline, total flying time, flying time in the 767, and their position (e.g., captain or first officer). In addition, a factor analysis was performed to determine if there were any underlying dimensions to the response to the 36 questions.

The pooled responses appear in Table 4 for each of the 36 questions.

Airline Differences A contingency table analysis was first performed to determine whether or not there any gross differences existed between airlines. The responses were pooled into a 3 X 3 matrix consisting of the three airlines and the three responses "agree/neither/disagree". There results were not significantly different from that expected by chance, thus returns were combined across airlines for later analyses.

Captains vs First Officers Each of the 36 questions was examined to determine if Captains and First Officers responded differently. This was done by constructing a 2 (captain/FO) X 2 (agree/disagree) contingency table for each of the 36 statements. There were 11 statements in which the captains and First officers agreed ($p > .80$, Table 5), and there was significant disagreement ($p < .05$) on two statements: the captains agreed (and the FOs disagreed) that the autoland capability enhances safety, and that "automation frees me of much of the routine, mechanical parts of flying so I can concentrate more on managing the flight".

Total Flying Time and 767 Flying Time An analysis was performed on the answers to the 36 opinion questions in order to determine the relationship between total flying time, 767 flying time, and captain/first officer differences and these opinion responses. This was done by performing a discriminant analysis to see if the three variables could discriminate between the two categories (agree/disagree) on each question. While there was some effect for a few statements (e.g., 767 time predicted agreement with the statement "I can find the exact location of important controls and switches without any hesitation"), in general, the percentage of correct classifications of responses on the basis of these three variables was always less than 70%, i.e., there seems to be almost no detectable relationship between the agree/disagree responses and the three variables.

Factor Analysis The responses to the 36 questions were subjected to a factor analysis (there were 96 complete responses for this purpose). An examination of the percent variance explained versus the number of factors showed no significant "knee" in the curve, but 8 factors explained slightly more than 60% of the variance. See Curry (1984) for more information on the factor analysis.

DISCUSSION

Pilot Acceptance of the New Technology

The Airplane in General

The pilots feel positively about the airplane. More than 86% agreed they "enjoy flying the 767 more than the older aircraft" (#11). In response to a statement (#34) about the enjoyment of hand flying, one pilot remarked "It's a sweetheart—tough to turn it over to automation!". This enthusiasm was also evident during the pilot interviews and the cockpit observations where the pilots also mentioned the aircraft performance (high climb rate and cruise altitudes) and the low fuel consumption.

The New Cockpit Technology

The pilots also seem accepting of the new cockpit technology, they choose to use it, and they find it helpful. Over 87% say they "like to use the new features of the 767 as much as possible"(#18), 79% "use the automatic devices a lot because I find them useful"(#10), although 31% also agreed to some degree that they "use automatic devices mainly because the company wants me to"(#35).

The items mentioned by the pilots are shown in Table 3. Particularly noteworthy is that the general capabilities of the AFDS, FMS/CDU and EICAS are mentioned, suggesting their general agreement with the functions and implementations. Specifically mentioned items, such at the map display and autothrottle, are also heavily used as indicated in the frequency of use table (in spite of their complaints about the implementation details of the autothrottle, Curry, 1984).

Workload

The pilot acceptance of the new cockpit technology, with respect to workload reduction, seems divided into two groups: those who say it *reduces* workload, and those who feel operating the devices *creates* a form of workload. This is reflected in the evenly divided responses to several questions: 47% agree and 36% disagree, that "Automation reduces overall workload" (#32); 53% agree and 37% disagree that "automation does not reduce overall workload, since there is more to keep watch over" #15; yet 79% agree that "I use the automatic devices a lot because I find them useful"(#10), regardless of any workload penalty. A workload issue for which there was a significant difference between captains and first officers seems based on their different roles: captains agreed more, on the average, and first officers disagreed more, on the average, that "Automation frees me of much of the routine, mechanical parts of flying so I can concentrate more on 'managing' the flight" (#24).

Skill Maintenance

Maintenance of flying skills was a concern of the pilots. This appeared in the questionnaires and in the pilot interviews. For example 87% agree that they "hand fly part of every trip to keep my skills up" (#14), and 80% agree that "pilots who overuse automation will see their flying skills suffer" (#18). Interestingly, this concern for other pilots did not always carry over to themselves because only 63% agreed that "I am concerned about a possible loss of my flying skills with too much automation" (#31).

The frequency of use table (Curry, 1984) shows that the pilots, in general, hand fly during transition and enroute climb (especially at the lower altitudes, as observed on line flights) and in the terminal area and final approach phases.

Equipment Reliability

Pilot opinion about the reliability of the equipment was measured by some of the attitude questions and roughly one quarter of the pilots expressed some concern. 20% of the pilots disagree with the statement "The new equipment is more reliable than the old" (#29) (45% agreed with the statement, and 35% neither agreed nor disagreed). Similarly, 27% agreed that they were "worried about sudden failures of the new devices like the FMS computer and the CRT displays" (#9), although the majority, 64%, disagreed with the statement; and 26% agreed that they "have serious concerns about the reliability of this new equipment", and again the majority disagreed (62%).

Features Disliked

There were few features or concepts that the pilots did not like, although there were features whose implementation, they felt, needed improvement.

FMC Response Delay **A large number of pilots felt that the response time for the Flight Management Computer was excessive. When a specific instance was mentioned, it usually involved complying with ATC requests while maneuvering in the terminal area. Although some of the pilots have learned that they can "type ahead" of the FMC, that is, push the appropriate buttons before the display requests the information, no one said they did this in the terminal area when rapid, accurate responses were required, perhaps because it has the potential for committing errors.**

Mechanical/Electrical Checklists **One of the participating carriers used a mechanical checklist, and the two others used cardboard checklists. Pilots of those two carriers felt some aid would be useful, especially as one pilot commented, it is difficult for a two man crew to get through a checklist without some form of interruption. Many of the pilots felt that having the checklist displayed on the EICAS would be beneficial. Perhaps so, but previous experiments (Rouse and Rouse, 1980) have found that simply transferring material to the CRT does not necessarily improve performance.**

Location of Circuit Breakers and Spare Bulbs **Several pilots commented on the inability to reach circuit breakers and spare bulbs while remaining in their seat. This appears to be a result of having to pull circuit breakers frequently during the early months of line operation to remove nuisance EICAS messages. The need to do this has been decreasing as system parameters are adjusted.**

Although the indicators have more than one bulb, one pilot reported having both bulbs in the landing gear indicator burned out. The cockpit design philosophy clashes with the reality of line operation at this point: should the pilot continue the landing without leaving his seat, or should he get up to replace the bulbs? Only more experience can answer this question.

Control Wheel Steering **This autopilot mode was rarely used by the pilots, and some said its use was discouraged during training. See Curry (1984) for further discussion of this topic.**

Points of Confusion and Surprise

Autothrottle-V/S-SPD Interactions

About 25% or the pilots reported experiencing some confusion, or seeing others become confused about the interaction of the autothrottles and autopilot. The source of this confusion seems to be twofold:

First, the thrust/elevator combination is a complicated interaction in any aircraft, and it recalls the seemingly endless debate about controlling speed/altitude with throttle/elevator. Obviously, both strategies are possible in climb and descent. (There is agreement in some regimes, such as constant altitude: elevator controls altitude, thrust controls speed.) When these functions are automated, then, confusion and surprise are likely to follow if the pilots are not aware of the modes actually in use.

The second proposed reason for the confusion of the autopilot/autothrottle interactions, is that this design has more features than previous systems. The autothrottle is almost always "armed"; in this state, it can become engaged, e.g., by engaging the SPD mode, even though it had been turned off with the throttle-mounted switches. Most pilots are used to autothrottles that can only be engaged by an autothrottle switch. The response to the questionnaires and the experience in line observation suggests that there is some uncertainty about the conditions that will allow the autothrottles to become engaged. In addition, the throttles seem to come out of idle during descent at times that the pilots feel are inappropriate.

Almost 10% of the pilots reported some discomfort with the speed syncronization at the time the Flight Level Change (FLCH) mode is engaged; FLCH is designed to climb at the existing IAS and climb thrust. The reason for the confusion seems to be that the SPD window shows a value at the time FLCH is engaged, but this value has no bearing on FLCH operation since the displayed speed automatically changes to the existing speed when FLCH is engaged. These pilots felt that FLCH should hold the speed displayed in the window, instead of the existing speed. Perhaps the confusion arises because the other numerical parameters on the mode control panel (altitude, heading, even speed itself) operate as selected, not held, values.

It is difficult, from the available data, to allocate the the autothrottle/autopilot confusion among the several possible sources: system design, system implementation, training, and lack of experience with the aircraft.

AFDS Turns "Wrong Way" or doesn't Capture

Nearly 20% of the pilots reported that at one time or another, the autopilot either turned the wrong way (usually on LOC intercept or passing over a waypoint), or did not capture the desired route or course. It is impossible from the reports received to attribute these occurences to a lack of system knowledge, incorrect programming of the system, or equipment malfunction. Even if the pilots could be contacted for more information, it would be difficult for them to recall all the pertinent details, and in addition, they may not know what caused the anomaly. (Some pilots, in their response to the question "have you ever been surprised by the automatics" answered in the affirmative, but said they never had the time to determine why).

The causes of reported "failure" of the FMS to capture a course are difficult to determine. It is true that several preconditions must be satisfied before capture will occur, and it was noted that not everyone was aware of these preconditions during the early phases of operation. Still, equipment malfunctions or idiosyncrocies cannot be ruled out as contributors to the reported instances.

Using The Wrong Control

Pilots report using the wrong control knob, especially the heading knob for the speed select knob, and vice versa. This seems to occur during the first few hours on the airplane, and disappears with exposure; no occurances were observed on the line trips.

Unselected Mode Changes

This phenomenon was reported by 12% of the pilots, with all but two reporting a change to vertical speed, and the others reporting a change to heading hold; both are the default modes of the autopilot. One incident was precipitated by such a change.

Training

Subject Material Grouping

Conversations with personnel involved in the transition training suggested that pilots felt the material fell naturally into three topics: aircraft systems, the Autopilot and Mode Control Panel, and the Flight Management System. In some sense, the same was true for the instructors and program developers. Both the pilots and instructors seemed more at home with the aircraft systems, and these were learned without any appreciable difficulty even though they sometimes contained more automation than previous systems, e.g., electrical source selection. Some pilots and instructors had previous experience with mode control panels. Instructors felt strongly that this previous experience made the transition easier for pilots.

The Flight Management System was entirely new to most instructors and pilots. Although some had prior experience with inertial navigation systems, the extensive capabilities of the FMS and its integrated nature were completely new to most individuals. The following comments received from two pilots reflect this view.

> "[The FMS/CDU] system is complex and so completely different."
>
> "I believe that the FMC was the most difficult to understand during ground school and the first few periods in the simulator. My classmates felt the same way."

FMS/CDU Training

When asked on the questionnaire what material they wanted more or less of in training, the strongest responses were requests for: more FMS and CDU training (in general); more "hands on" experience and training with the FMS/CDU; more line-oriented CDU exercises; and less non-operational CDU material. These comments were confirmed by several line training pilots, who, in the early phases, felt that the pilots arrived for line training with less than desirable knowledge and skills about the FMS/CDU.

The difficulties of conducting the FMS/CDU training seemed to have come from several sources. First, there were many new concepts for the pilots to learn, e.g., navigating from autotuned radios, not from a single radio. Second, although it is beyond the scope of this study to identify the conceptually difficult aspects of the system, the organization of the information, and the naming conventions seemed to cause problems for people. Third, and perhaps most important, there was no training device that (from the pilots' view) was an adequate simulation of the real FMS/CDU; see the comments below on Computer Aided Instruction.

Relevant Material

It can be seen from the responses that many of the pilots wished they had had more "realistic" or line-oriented material in their FMS/CDU exercises and less material on features that were not operational. This latter request seems to have arisen from the scheduled versus actual introduction of equipment capabilities (Curry, 1984).

In addition to the material they received that they did not need, the pilots also felt that they did not receive material they could have used. In the case of the FMS/CDU, pilots revealed in interviews that they did not know how to deal with tasks such as crossing restrictions until after their line training. Although one can argue that these functions would have been covered by the VNAV system, pilots were not given an interim method and sometimes did not receive the material in line training. Another item mentioned in the interviews, and the questionnaires, was a last minute change in approach assigned by ATC; removing old information seemed to be as much of a problem as selecting the new approach from the menu.

Computer Concepts

Two of the questionnaire respondents asked for some instruction on computer concepts.

> "Ground school should not teach just function of the CDU/computers, but a philosophy of computer applications and programming as applicable to our aircraft. This was done when the [new jet turbine technology] B-707 was introduced in 1958. Now that everyone is jet oriented, this is not necessary. So today, the computer is new and should be taught until everyone has the 'idea'".
>
> "For those of us with no computer literacy (buzz word) a 10 minute dissertation on computer functioning would help. Actually, just the thought that the damn thing only does what it is told would save some errors."

One pilot suggested an even broader scope.

> "From what I've seen so far, we could use a bit more emphasis on the 'background' of some of the automatics to better able a crew to understand what's happening or not happening when things don't go as programmed..."

This type of training would certainly be consistent with the idea of creating a "schema" or framework about computers or automation, into which detailed information would more easily be assimilated.

CONCLUSIONS

The data reported in this paper were taken from the introductory experience of the B-767, but it is felt that similar, if not identical, results would be obtained with any other new cockpit technology aircraft, i.e., the A310. The following conclusions have been drawn from the information collected thus far:

- A large majority of the pilots enjoy flying the B-767 more than the older airplanes.
- The pilots accept the new cockpit technology, and they choose to use it because they find it useful.
- The pilots are aware of the possible loss of flying skill with the presence of automation, and they hand fly (usually with flight director) to prevent this loss. There is no evidence of loss of skills from the data collected in this study.
- The primary points of confusion or suprise were autothrottle/autopilot interactions; the autopilot turning the "wrong way" or not capturing the course; and achieving desired results with the FMS/CDU.
- The pilots felt training for the FMS/CDU (Flight Management System/Control Display Unit) could be improved, and they especially wanted more "hands on" experience. More training on the mode control panel, and more hand flying were also mentioned.

REFERENCES

Curry, R.E., The introduction of new cockpit technology: a human factors study, NASA TM (in preparation), 1984

Rouse, S.H. and Rouse, W.B., Computer based manuals for procedural information, *IEEE Trans. Syst., Man, and Cybern.*, SMC-10, pp506,510, 1980

Wiener, E.L., and Curry, R.E., Flightdeck automation: promises and problems, *Ergonomics*, *23*, 995-1011, 1980

Table 1. Statements Rated by Pilots

1. I can fly the airplane as smoothly and safely by hand as with automation.
2. Younger pilots catch on to automation faster than older ones.
3. Flying today is more challenging than ever.
4. The FMS/CDU is easy to use in normal line flying.
5. I think they've gone too far with automation.
6. Autoland capability definitely enhances safety.
7. I spend more time setting up and managing the automatics (such as the FMS/CDU) than I would hand flying or using the old style autopilots.
8. I like to use the new features of the 767 as much as possible.
9. I am worried about sudden failures of the new devices like the FMS Computer and the CRT displays.
10. I use automatic devices a lot because I find them useful.
11. I enjoy flying the 767 more than the older aircraft.
12. I always know what mode the Autopilot/Flight Director is in.
13. I can fly as efficiently as the FMS without its help.
14. I hand fly part of every trip to keep my skills up.
15. Automation does not reduce workload, since there is more to keep watch over.
16. I can find the exact location of important controls and switches without any hesitation.
17. Automation is the thing that is going to turn my company around and make it profitable again.
18. Pilots who overuse automation will see their skills suffer.
19. The ADI and EHSI displays are always legible and easy to read.
20. I am favorable toward automation in the cockpit - the more the better.
21. Flying the 767 is definitely easier than flying other aircraft.
22. Setting piloting priorities with this new cockpit technology is no more difficult than in our other airplanes.
23. We should have full autothrottles on all the company's aircraft.
24. Automation frees me of much of the routine, mechanical parts of flying so I can concentrate more on "managing" the flight.
25. I have serious concerns about the reliability of this new equipment.
26. Sometimes what the automatics do or don't do takes me by surprise.
27. It is easier to cross-check the other pilot in the 767 than in our other airplanes.
28. Too much automation can be dangerous.
29. The new equipment is more reliable than the old.
30. It is important to me to fly the most modern plane in the company's fleet.
31. I am concerned about a possible loss of my flying skills with too much automation.
32. Automation reduces overall workload.
33. I always feel I am ahead of the airplane.
34. Hand flying is the part of the trip I enjoy most.
35. I use automatic devices mainly because the company wants me to.
36. The FMS/CDU requires little or no in-flight button-pushing below FL180.

Table 2 Pilot Statistics

	Captain	F/O	Total Time(hours)			767 Time(hours)		
			Minimum	Median	Maximum	Minimum	Median	Maximum
Airline A	15	7	8000	14000	23150	17	60	300
Airline B	16	12	8500	12000	24000	20	113	300
Airline C	30	22	4200	15500	25000	5	104	250
All Pilots	61	41	4200	13500	25000	5	100	300

Table 3. Specific items mentioned on questionaires

Frequency

Total
102

FEATURES LIKED

AFDS

20	Autothrottle Concept/Speed Control
14	AFDS Capabilities
10	Takeoff Mode and/or EEC
8	Reduced Workload
6	Altitude Capture/Select

EFIS

42	Display and clarity of information
22	Map display
7	Green Altitude Arc
5	Wind Vector
4	ADI Mode Annunciation
2	Ground speed display

FMS/CDU

48	System capabilities
2	Route display

EICAS

35	Quality & Quantity of information
6	Engine limits & numbers
3	Monitoring capabilities

FEATURES MISSING OR NOT LIKED

20	FMC response delay
7	Want electrical/mechanical checklists
7	Circuit breakers and spare bulbs not within reach

Table 3. Specific items mentioned on questionaires (cont.)

POINTS OF CONFUSION OR "SURPRISE"

25	Autothrottle-V/S-SPD Interaction
20	AFDS turns "wrong way" or does not engage
19	Using wrong control (especially HDG/SPD)
12	Unselected mode change (10 to V/S, 2 to HDG HLD)
11	Removing route discontinuities and extra information
11	Track/heading on map display
9	Speed sync at FLCH engagement
7	Early altitude capture at high climb rate
7	AFDS-MCP mode (general)
6	FMS/CDU useage (general)
6	Simultaneous Speed brakes & landing flaps
5	Changing approaches on FMS/CDU close-in
3	No aural trim indication
3	Holding with FMS/CDU
3	Map drift
2	Use of J routes in FMS/CDU
2	High bank angles at LOC capture
2	Defining waypoints from station
1	Unselected level-off at FL180

TRAINING

4	Satisfactory as is

More:

25	FMS/CDU
22	"Hands on" CDU experience
12	Hand flying
8	AFDS-MCP training
7	Practical, line-oriented CDU exercises
6	Aircraft systems
3	Single engine simulator experience

Less:

10	Computer aided instruction
7	3 man simulator
3	non-operational FMS material
2	Phase-of-flight presentation

Table 4. Pilot Opinion Summary (% responses in each category)

QUESTION NUMBER	STRONGLY AGREE	SLIGHTLY AGREE	NEITHER AGREE NOR DISAGREE	SLIGHTLY DISAGREE	STRONGLY DISAGREE
1	28	31	12	25	5
2	12	37	28	18	6
3	37	29	14	17	3
4	38	35	5	19	3
5	3	17	18	27	35
6	26	36	17	15	6
7	30	35	7	15	14
8	54	36	6	4	0
9	11	16	10	28	36
10	39	40	16	6	0
11	62	24	7	7	0
12	29	32	8	28	3
13	3	17	18	40	23
14	63	24	4	7	2
15	22	31	10	23	14
16	29	29	10	31	2
17	6	15	39	18	21
18	48	32	6	12	3
19	51	28	5	13	3
20	15	44	17	17	6
21	13	33	24	25	6
22	19	32	10	32	8
23	16	21	41	15	8
24	19	42	16	19	5
25	4	22	13	30	32
26	10	52	8	22	8
27	11	26	30	27	7
28	11	34	29	17	10
29	13	32	35	21	0
30	16	28	33	16	8
31	24	39	8	16	13
32	18	29	17	31	5
33	21	40	10	28	1
34	22	38	25	10	6
35	6	25	27	30	13
36	3	13	6	24	55

Table 5 Contingency Table Comparisons of Captains vs First Officers and their Response to the 36 Statements

Statements on which there was agreement ($p > 0.80$)

Probability	Statement number	Statement
.85	1	I can fly the airplane as smoothly and safely by hand as with automation.
.99	4	The FMS/CDU is easy to use in normal line flying.
.87	10	I use automatic devices a lot because I find them useful.
.88	12	I always know what mode the Autopilot/Flight Director is in.
.85	13	I can fly as efficiently as the FMS without its help.
.91	19	The ADI and EHSI displays are always legible and easy to read.
.90	21	Flying the 767 is definitely easier than flying other aircraft.
1.00	22	Setting piloting priorities with this new cockpit technology is no more difficult than in our other airplanes.
.87	23	We should have full autothrottles on all the company's aircraft.
.85	32	Automation reduces overall workload.
.84	34	Hand flying is the part of the trip I enjoy most.

Statements on which there was disagreement ($p < .05$)

Probability	Statement number	Statement	Reasons
.047	6	Autoland capability definitely enhances safety.	Captains agree more, FOs disagree more
.043	24	Automation frees me of much of the routine, mechanical parts of flying so I can concentrate more on "managing" the flight.	Captains agree more, FOs disagree more

N85
14555

UNCLAS

N85 14555

D20

Crew Communication as a Factor in Aviation Accidents

by Joseph Goguen[1], Charlotte Linde[2] and Miles Murphy[3]

1. INTRODUCTION

The basic motivation for the research reported here is to reduce the incidence of those air transport accidents caused wholly or in part by problems in crew communication and coordination. A major objective is to determine those communication patterns which actually are most effective in specific situations; this requires developing methods for assessing the effectiveness of crew communication patterns. It is hoped that these results will lead to the development of new methods for training crews to communicate more effectively, and will provide guidelines for the design of aviation procedures and equipment.

The two main contributions of this study are a set of validated hypotheses about air crew communication patterns, and the development of a novel methodology for formulating and testing such hypotheses. Transcripts from eight commercial aviation accidents were used as data. Section 3 below gives a precise treatment of each hypothesis, while Section 2 presents the procedures used, including definitions for the variables occurring in the hypotheses. The following list informally presents the result of testing each hypothesis, together with its relevance for aviation safety:

1. Speech acts to superiors are more mitigated, i.e., the speech of subordinates is more tentative and indirect than the speech of superiors. This indicates a relationship between the social hierarchy and the form of cockpit discourse, and provides a foundation for later hypotheses concerning the effects of excessive mitigation.
2. Speech acts are less mitigated in crew recognized emergencies, i.e., when crew members (including the captain) know that they are in a emergency situation, their speech is less tentative and indirect. This indicates that crew members are able to vary their use of mitigation depending on their perception of the situation, and suggests both that experienced crews feel that mitigation is inappropriate in an emergency and that the level of mitigation may be trainable.
3. Speech acts are less mitigated during crew recognized problems. This hypothesis is similar to the previous one, stating that when crew members know that they are in a problem situation, their speech is less tentative and indirect.
4. Captains and subordinates differ in frequency of planning and explaining. This

---

[1] Structural Semantics, P.O.B 707, Palo Alto CA 94302, and SRI International, Menlo Park CA 94025.

[2] Structural Semantics, P.O.B. 707, Palo Alto 94302.

[3] NASA, Ames Research Center, Moffett Field CA 94035.

PRECEDING PAGE BLANK NOT FILMED

hypothesis probes, in an indirect way, possible inhibitory effects of the social hierarchy on contributions by subordinates. Test results suggest that captains may plan and explain more than subordinates.

5. Planning and explanation are less common in crew recognized emergencies. This hypothesis represents the intuition that when crew members know they face an emergency, they will do less planning and explaining of possible courses of action. Clearly an emergency calls for immediate action, but it is still possible that more planning and explanation would be useful in some emergency situations.
6. Planning and explanation are more common during crew recognized problems. This hypothesis states that when crew members are aware that they are in a problem situation, they do more planning and explaining. This result indicates that crew members do indeed plan and explain in appropriate situations, those where the original flight plan is no longer adequate.
7. Topic failed speech acts are more mitigated. This hypothesis tests the idea that excessive mitigation can have undesirable consequences, specifically that a new topic is less likely to be picked up by other crew members if the speech act in which it is introduced is excessively mitigated. The result suggests that the frequent situation of a subordinate failing to get a correct point accepted might be improved by training in linguistic directness.
8. Unratified draft orders are more mitigated. This hypothesis tests the idea that when a crew member proposes a suggestion to the captain, the more indirect and tentative the suggestion, the less likely the captain is to ratify it. Like the preceding hypothesis, this suggests the possible value of training in linguistic directness.

These results show that crew communication patterns are frequently present in accident situations, and suggest that they may have a significant effect on aviation safety.

This methodology is novel in its use of linguistic investigation of how crews actually talk, using aviation accident transcripts as data, yielding an empirically grounded formal description of communication patterns in the cockpit. An ongoing study uses the same methodology and hypotheses with audio and video recordings of sixteen full mission simulations as data [Murphy *et al.* 84].

2. METHOD

This section discusses data acquisition, the theoretical concepts which define the variables used in the hypotheses, and the variables themselves.

2.1 Sampling Procedures

There are three main stages to the sampling process: (1) the production of accident transcripts, (2) the selection of transcripts, and (3) coding the selected transcripts. The sample space that results from these procedures consists of a large number of speech acts, rather than, for example, a small number of transcripts or of crew members. This choice seems well suited to studying relationships between linguistic behavior and features of the cockpit situation. On the other hand, accident transcript data is less suitable for studying individual differences in the behavior of crews or crew members. This is the case because these transcripts do not provide a sample of crews tested in a single standard situation, but rather show single crews in a variety of unique situations. We will not describe the production of accident transcripts here, except to note that this is an "unobtrusive" procedure, in the sense that the collection of this data has nothing to do with the researchers who later analyze it.

2.1.1 Transcript Selection Criteria

The transcript selection criteria were developed using categories and analyses in [Murphy 80]:

1. The transcript contains a critical segment. A **critical segment** is a portion of transcript containing observable degradation or failure of crew coordination which is actually or potentially critical to the completion of the flight.
2. The entire situation of interest must not be significantly longer than 30 minutes (since the maximum length of the tape is 30 minutes).
3. There must be sufficient background information to permit understanding all relevant aspects of the situation (e.g., in the NTSB report).
4. The language of the transcript should be suitable for analysis. In particular, there should be enough talk to permit analysis, and all the conversation should be in English, since we are not focussing on cross-linguistic problems.
5. There should be sufficient interest and agreement in the aviation community to support further investigation.
6. All other things being equal, more recent transcripts are preferred. (Note that this criterion also plays a major role in determining whether or not criterion 5 is satisfied; older flights are of lesser interest since the procedures and equipment are more likely to have been superseded.)
7. If possible, the set of transcripts should include all flight segments -- taxi, takeoff, climb, cruise, approach and land.

Eleven potentially suitable transcripts were preselected at NASA using criteria 1 and 5, and 6 and 7 whenever possible:

1. United Airlines/Portland/78;
2. Eastern Airlines/Miami/72;
3. Northwest Orient Airlines/Thiells, New York/74;
4. Allegheny Airlines/Rochester/78;
5. World Airlines/Cold Bay, Alaska/73;
6. Texas International Airlines/Mena, Arkansas/73;
7. Pan American Airlines/Bali/74;
8. Air Florida/Washington, D.C./82;
9. Southern Airways/New Hope, Georgia/77;
10. PSA/San Diego/78; and

11. Pan American Airlines/Teneriffe/77.

These eleven were examined in detail. The first eight transcripts of this set were found suitable for inclusion in the sample. The last three are unsuitable for the following reasons:

1. Southern/New Hope. Several of the major contributing events occur before the beginning of the tape, and indeed, before departure, i.e. the company's failure to provide up-to-date severe weather information, and the crew's "lack of significant attempt to seek information on current flight conditions" (NTSB report, p. 33).
2. PSA/San Diego. The NTSB report on this accident mentions the possibility that there were two small planes in the vicintiy of the PSA plane, rather than just one as both the crew and ground control appear to have believed. This makes it impossible to determine accurately to which plane the PSA crew and ground control were referring at any given time.
3. Pan Am-KLM/Teneriffe. Unlike the other accidents considered, the cause of this accident appears to lie in failure of communication with the tower, rather than in crew coordination. Futhermore, some of the communication problems appear to arise from the fact that three different languages are involved -- English, Spanish, and Dutch.

2.1.2 Data Coding

Each of the 1725 speech acts in the sample space was coded according to twelve categories: speaker; addressee(s); speech act type; discourse type; new or old topic; topic success or failure; draft order; ratification; mitigation level; crew recognized emergency; crew recognized problem; and operational relevance. Most of these variables depend upon linguistic theories that are described in the following subsection. For details of the coding procedures, see [Structural Semantics 83].

2.1.3 Use of Hypothesis Formulation and Test Transcripts

Two of the eight transcripts, chosen for the interest of their language and situation, were used to develop hypotheses which illuminate the basic structure of crew communications. We call these two transcripts, United/Portland/78 and Texas/Mena/73, the **hypothesis formulation group**. The remaining six transcripts were used to test the hypothesis; we call these transcripts the **test group**.

The six transcripts from the test group contain altogether 480 operationally relevant speech acts, and the two hypothesis formulation transcripts contain 399. Each hypothesis selects as a dataset for testing a subset of the 399 speech acts of the hypothesis formulation group and a subset of the 480 speech acts of the test group.

Each hypothesis is first tested on speech acts from the six transcripts of the test group. It is then tested on the speech acts from the two hypothesis formulation transcripts. Speech acts from these two groups are pooled when possible to yield a larger sample for a stronger test of the hypotheses. However, pooling is justified *only* if it is possible to avoid the methodological bias that results from testing hypotheses on the data from which they were formulated. If the hypothesis is accepted on the basis of data from the six test transcripts and/or is rejected for data from the two hypothesis formulation transcripts, then the two datasets can be combined. The purpose of this division is to reduce the probability that the obtained results are in actuality due to the effects of some uncontrolled variable.

2.2 Linguistic Theories

In order to provide an adequate description of cockpit communication, we have adapted or created a number of linguistic theories. These include speech act theory, and formal theories for the discourse types of planning, explanation, and command and control. These theories support the linguistic variables used in our hypotheses. The variables include: mitigation/aggravation level, crew recognized emergency, crew recognized problem, operational relevance, topic success or failure, and draft order and ratification. We now turn to a brief dicussion of these underlying linguistic theories.

2.2.1 Speech Act Theory

Speech act theory focusses on the operational aspect of language -- how a particular sentence achieves some effect in the world. We call this the **social force** of the speech act. The fundamental insight of speech act theory is that some sentences, such as (1), *describe* or *report* a state of the world, while other sentences, such as (2), *create* a state of the world.

(1) There's a thunderstorm ahead.
(2) I declare this bridge open.

These examples express their social force directly. However, there are also speech acts which express their most probable social force indirectly, by using a linguistic form which is not to be interpreted literally. For example[4]

(3) CAM-1 What I need is the wind, really
(1755:13)

is literally an **expressive**, in which the captain expresses his psychological state of "needing" information about the wind. However, given the context in which it was spoken, its social force might be given as the directive

---

[4] Examples from the United/Portland/78 transcript are indicated by giving the time of the utterance below it in parentheses. This transcript is used in our examples because of its relevance to our research topic and its familiarity to the aviation community.

(4) Give me the wind.

The basic question about indirect speech acts is how can it happen that one speech act gets interpreted as another. To answer this, speech act theory [Searle 79] uses **felicity conditions**, which are conditions that must be satisfied in order for a speech act of a given kind to be uttered "felicitously" (also termed "non-defectively"). These conditions include preparatory conditions, propositional content conditions, sincerity conditions, an essential condition, and possibly some others. **Preparatory conditions** cover what must be satisfied before the utterance is made; for example, for an order, that the speaker must have appropriate authority over the addressee, and that the addresee is able to perform the act; or for a promise, that it is not obvious that what is promised would otherwise occur. **Propositional content conditions** express constraints on the propositional content; for example, for a promise, that it express a future act by the speaker. **Sincerity conditions** concern the speaker's internal states, including his intentions. For example, in a request that the addressee perform an act A, the speaker should really want the addressee to do A. The **essential condition** defines the desired effect of the speech act upon the addressee.

Preparatory:	**Addressee is able to perform act A**
Propositional Content:	**Speaker predicates a future act A of the addressee**
Sincerity:	**Speaker wants the addressee to do act A**
Essential:	**Utterance counts as an attempt by the speaker to get the addressee to do act A**

Figure 1: Felicity Conditions for Directives

The most obvious way to accomplish a speech act indirectly is to make reference to one of its felicity conditions. For example, a sincerity condition for a request that the addressee make a report is that the speaker should really want to know the contents of this report. This gives us an explanation of how (3) can indirectly convey (4). Figure 1 gives a list of felicity conditions for directives, a class which includes orders and requests; Figure 2 gives a list of "generalizations" for the indirect accomplishment of directives. (Both figures are adapted from [Searle 79].)

---

1. **Preparatory Condition**. Speaker can make an indirect directive to do act A either by asking whether a preparatory condition concerning the addressee's ability to do A holds, or by stating that it does hold.

2. **Propositional Content**. Speaker can make an indirect directive by asking whether the propositional content condition holds or by stating that it does hold.

3. **Sincerity Condition**. Speaker can make an indirect directive by stating that the sincerity condition holds, but not by asking whether it holds.

4. **Essential Condition**. Speaker can make an indirect directive to do an act A either by stating that there are good or overriding reasons for doing A,or by asking whether such reasons exist, except where the reason is that the addressee wishes to do A, in which case the speaker can only ask whether the addressee wishes to do A, but can not assert that he does.

Figure 2: Strategies for Indirect Directives

There is a very large literature on indirect speech acts in the fields of linguistics, philosophy of language, artificial intelligence, and psychology, e.g. [Searle 79, Gordon & Lakoff 71, Gazdar 79, Labov & Fanshel 77]. Our discussion summarizes the approach of [Searle 79], which underlies most other approaches. Speech act theory also provides a taxonomy of possible types of speech act. We have modified this taxonomy to provide an inclusive listing of the speech acts found in cockpit communication. These are: **requests**, including orders, requests, suggestions and questions; **reports**; **declarations**; and **acknowledgements**. The set of all speech acts in the 8 transcripts selected in Section 2.1 constitute the basic sample on which our hypotheses were tested.

2.2.2 The Discourse Unit

Although speech act theory is of great value to the study of crew communication, our account would be quite incomplete if it remained at the level of the simple sentence. It is necessary to study larger units as well. The larger unit of language that we have found appropriate for this study is called the discourse unit. A **discourse unit** is a segment of spoken language, longer than a single sentence, having initial and final boundaries that are socially recognizable, and having a formally definable internal structure. (This definition generalizes the criteria given by [Labov 72] for the narrative of personal experience.) Discourse types that have been studied include the narrative, the spatial desciption [Linde 74, Linde & Labov 75], the joke [Sachs 74], small group planning [Linde & Goguen 78], and explanation [Goguen, Weiner & Linde 83]. There are a number of points to be made about the definition of discourse unit:

1. Level of Unit. In the linguistic hierarchy, the discourse unit is immediately above the sentence, and hence is composed of sentences.
2. Socially Recognized Boundaries. The discourse unit has boundaries which are recognized as such by the participants in the conversation. These boundaries are often recognized through their stereotyped form; for example, **They lived happily ever after** as the end of a fairy tale, **It seems there was a**. . . as the beginning of a joke, **And that was it.** as the end of a narrative. Or they may be recognized as encoding a certain type of semantic information; for example, an abstract of a story, summarizing its point, can serve as an initial boundary.
3. Formally Definable Internal Structure. Labov has given an account of the structure of narrative which is, in effect, a phrase structure grammar [Labov 72]. Plans and reasoning have been described using a transformational grammar in which the transformations mirror the real-time additions, deletions, and modifications made by speakers [Goguen, Weiner & Linde 83, Linde & Goguen 78]. A **discourse type** is a class of discourse units having internal structure in conformity with the same set of rules.

We have found that the most important discourse types in the study of crew communication are planning, reasoning, and the command and control speech act chain. We have also found instances of narrative and pseudonarrative, but since they are used only in non-operationally relevant ways, we do not consider them here. The following subsections discuss the three operationally relevant discourse types in some detail.

2.2.3 Planning

This research focusses on planning as a linguistic activity carried on by a group rather than as an individual mental activity. The linguistic study of small group planning [Linde & Goguen 78] has shown that the language used to accomplish planning is a discourse type, since it has an initial boundary, consisting of the statement of the goal which the planning is intended to accomplish; it has a final boundary, which may consist of the group's evaluation of the probable effects of the plan, or of their acceptance or rejection of it; and it has a precise internal structure, consisting of members' proposals to add new subplans, and to modify or replace parts of the plan previously proposed by others.

Formally, the internal structure of a planning discourse unit is described as a sequence of **transformations** on the plan being formed by the group. In planning, these transformations represent the real-time effects of proposals by members to add, delete, or modify plan parts. Similarly, the relations of logical subordination that hold among the plan parts are represented by a **tree structure**. Figures 3 and 4 show a plan from the United/Portland/1978 accident. Its major goal, stated by the first officer, is to `call out the equipment`, and his plan for this is to `have the company call`. This PLAN/GOAL relationship is indicated in Figure 3. In Figure 4, the captain replaces the first officer's plan with a plan to `call dispatch in San Francisco`. In Figure 5, he adds a node indicating that `maintenance down there will handle it that way`.

```
CAM-2 He's going to have                      PLAN/GOAL
      the company call                         /   \
      out the equipment?                      /     \
                                       have the     call out
                                    company call    the equipment
```

Figure 3: A GOAL/PLAN Node

```
CAM-1 We'll call dispatch                      GOAL/PLAN
      in San Francisco                          /   \
                                               /     \
                                      call out     ACTOR/SAY/TO
                                 the equipment      /    |   \
                                                   /     |    \
                                                 we    call  dispatch
                                                             in San
                                                             Francisco
```

Figure 4: Addition of an ACTOR/SAY/TO Node

2.2.4 Explanation

We do not use the term **explanation** to refer to segments of discourse that serve the function of explaining something; rather, explanation is a discourse type, having similar structural properties and expressible with similar formalism, as planning. Informally, an explanation is a discourse unit consisting of a statement to be demonstrated, and a structure of supporting reasons, which often have further embedded relationships of subordination. This kind of discourse occurs, for example, in social contexts where a

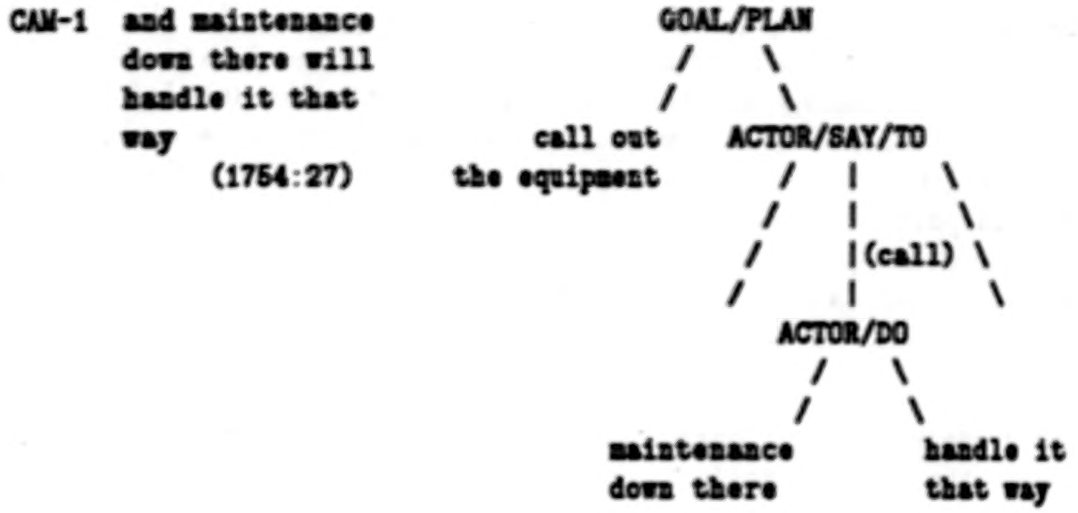


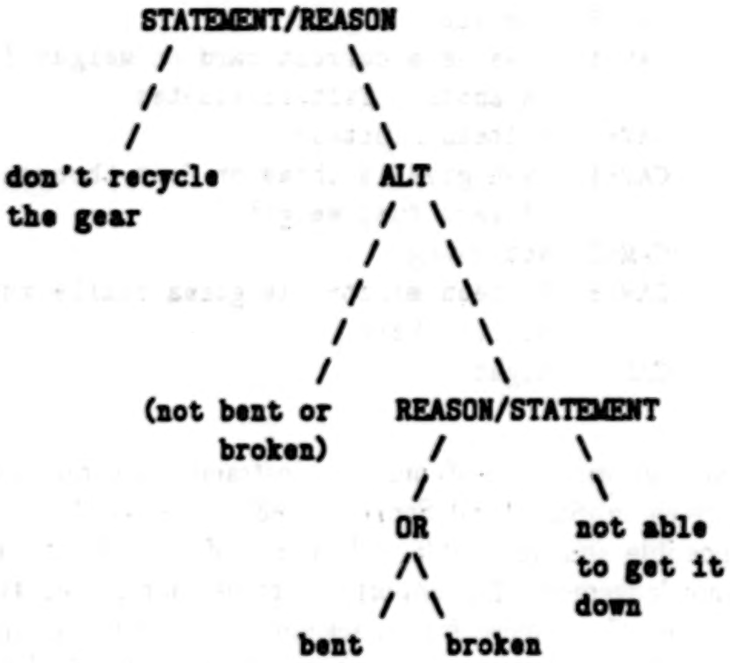
Figure 5: Addition of an ACTOR/DO Node

single person attempts to justify to an addressee actions he has already performed, or will perform later. Figure 6 shows an analysis of a simple explanation in which the flight engineer reports his justification of the decision not to recycle the landing gear.

The most important relationship of subordination in explanation is indicated by STATEMENT/REASON nodes. In Figure 6, the main STATEMENT is **Don't recycle the gear**. Everything that follows is a REASON supporting this. The ALT node represents the speaker's postulation of two alternate worlds, which differ by whether or not the landing gear is broken. This ALT node is established by the **underlined** portion of the text ...**we're reluctant to recycle the gear for fear something is bent or broken**. The phrase for fear indicates indicates both the uncertainty about whether the gear is bent, and the decision to treat the alternate world in which it is bent as the one on which attention is focussed.

2.2.5 Command and Control Discourse

The command and control speech act chain is the basic discourse type for command and control in the cockpit. A **speech act chain** is a sequence of speech acts, each of having the same major propositional content. Command and control chains may also include the other discourse types characteristic of operationally relevant cockpit communication, planning and explanation. (5) is a typical speech act chain. Its component speech acts include requests, reports, explanations and acknowledgements, all concerning the major topic "fuel weight."


STATEMENT/REASON
don't recycle the gear
ALT
(not bent or broken)
REASON/STATEMENT
OR
not able to get it down
bent
broken

... and I said we're reluctant to recycle the gear for fear something is bent or broken, and we won't be able to get it down (1751:16)

Figure 6: An Explanation Tree

(5a)	**CAM-1**	**Hey Frostie**
(5b)	**CAM-3**	**Yes sir**
(5c)	**CAM-1**	**Give us a current card on weight figure in another fifteen minutes**
(5d)	**CAM-3**	**Fifteen minutes?**
(5e)	**CAM-1**	**Yeah give us three or four thousand pounds on top of zero fuel weight**
(5f)	**CAM-3**	**Not enough**
(5g)	**CAM-3**	**Fifteen minutes is gonna really run us low on fuel here**
(5h)	**CAM-?**	**Right**

(1750:16)

Details of a formal grammar specifying the constraints on command and control speech act chains are given in [Structural Semantics 83]. One use for such a grammar is to indicate some possible and impossible embeddings of social force. For example, we will not find an acknowledgement of a support of a request for an act. However, we may find an acknowledgement of a request for an act and a request for a support of a request for an act. We hypothesize that correct command and control chains describe optimal patterns of communication in the cockpit, particularly in emergency situations. Although this hypothesis can not be tested using accident data, it could be tested with simulator experiments by training crews to use strict command and control form, then measuring flight performance, and comparing it with performance of a control group not so trained.

2.3 Variables Used in the Hypotheses

The discourse types discussed above provide many of the coding categories used in the precise formulation of the hypotheses, including the division of the transcript into speech acts, the inclusion of a speech act in a command and control chain, or in planning or explanation. We now discuss the remaining variables required to code the data for hypothesis testing.

2.3.1 Crew Recognized Emergency

Crew Recognized Emergency (CRE) is a social, rather than a legal or factual category. The beginning of a crew recognized emergency is defined as the first point at which the entire crew begins to attend to that situation which led directly to the accident. There are several remarks to be made about this definition:

1. In order to identify the situation which led to the accident, we rely upon informed and documented opinion in the aviation community. In practice, this means that we rely on NTSB accident reports, but in disputed cases, it would be possible to make use of minority reports, other published materials, or oral reports from members of the aviation community.

2. The definition requires that the entire crew attend to the situation. It may happen that individual crew members attend to the situation that led to the accident long before the CRE point, and may even have attempted to bring it to the attention of the rest of the crew. However, it is group attention that is required here. Note that in practice, this means the attention of the captain, since in the command and control situation, the captain has the authority to direct the attention of the crew to any situation which he considers to be threatening, while other crew members may suggest but can not compel such attention.
3. In some accidents there may never be a CRE. These are cases in which the crew never attends to those situations that caused the accident.

Note that a captain's official Mayday declaration does not serve to identify this point, since this declaration often appears quite late, considerably after the point at which the crew begins to act as if they were in an emergency situation. In fact, Mayday is a legal category, specifying a situation in which there is "immediate danger to equipment and personnel."

2.3.2 Crew Recognized Problem

In addition to Crew Recognized Emergency, we also use the notion of **Crew Recognized Problem** (CRP). This is a situation recognized by the crew as potentially dangerous and not a normal part of flight operations. It could be an actual problem, or some situation which is off-nominal, surprising, or not expected. Like CRE, Crew Recognized Problem is not defined by the actual onset of the situation as given by the system data readout, but by the point at which the crew as a whole first attends to that situation.

2.3.3 Operational Relevance

A very pervasive distinction, entering into many of our definitions and all of our hypotheses, is whether or not a given speech act is operationally relevant. Operational relevance means that the speech act is *directly* involved with the achievement of the mission. Thus, a request for a snack is not operationally relevant, even though it might have some effect on the state of a crew member, and hence an indirect effect on successful mission completion. This distinction has been introduced because there are certain phenomena which are potentially of great importance in operationally relevant discourse, but have no serious consequence in non-operationally relevant segments. For example, if a speaker introduces an operationally relevant topic and other crew members do not continue this topic, the consequences can be quite serious. However, failure of a non-operationally relevant topic is of much less concern.

2.3.4 Mitigation/Aggravation

The **mitigation/aggravation** scale provides one way of assessing the assertiveness of speech acts. For example, (6) is direct, (7) is mitigated, (8) is highly mitigated, and (9) is aggravated.

(6) `Close the window.`
(7) `Would you close the window?`
(8) `Please, would you mind closing the window?`
(9) `Listen, close that damn window right now.`

Mitigation softens the possible offense that an utterance might give. Our results show that mitigation is very important for cockpit communication, since the greater the degree of mitigation, the more likely it is that a given utterance will fail to accomplish its effect, and speech acts by subordinates are more mitigated than those of superiors (Section 3.1 gives a more precise discussion of these findings).

It should be noted that mitigation and aggravation are linguistic categories, not psychological ones. Thus, when a speaker uses an aggravated form, we can not directly draw any conclusions about his psychological state at the moment, nor about his personality characteristics, although a speaker's long-term profile of using mitigation/aggravation in different contexts is probably related to his personality characteristics. Use of few mitigation strategies, or of many aggravation strategies is one way of behaving assertively; there are, of course, many others. There are many devices that function to mitigate: questions are more mitigating than imperatives; modal auxiliaries such as `would`, `might` and `could` are more mitigating than simple verbs; markers of request for agreement such as `right` and `OK` are mitigating. Moreover, indirect speech acts (see Section 2.2.1) are more mitigating than direct ones. (See [Structural Semantics 83] for a review of the unified theory of mitigation given by [Brown and Levinson 79].)

2.3.5 Scale of Mitigation/Aggravation

Several of the hypotheses suggested in this report require discriminating degrees in a scale of mitigation and aggravation. The degrees of this scale correspond to the sense felt by native speakers of a language that some sentences are more polite or more indirect than others. The validity of this scale has been established by checking the judgement of linguistic analysts against the judgements of members of the aviation community. We have found that four degrees of mitigation/aggravation are the most that native speakers can reliably discriminate. This scale has a midpoint of zero, representing a direct, unmitigated utterance. There are two degrees of mitigation -- low and high. There is only one degree of aggravation, corresponding to the facts that aggravation is much rarer than mitigation and that there are fewer strategies for effecting aggravation than for effecting mitigation.

Scale validation was established by a reliability experiment. The stimuli consisted of 31 reports and requests, chosen randomly from the six transcripts. The scale of mitigation/aggravation tested had the following four levels: Aggravated; Direct; Low Mitigation; and High Mitigation. The experimental subjects consisted of six commercial airline professionals, including two captains, three first officers, and one flight engineer. Before being asked to score the speech acts, they were given pre-test training in the meaning of the categories used: A previously prepared explanation of the notion of mitigation was read to the subjects. They were then given some sample speech acts (in written form) to rate, and these examples were discussed with the group by one of the analysts. Finally, they were given the written stimuli to score.

The criterion which is generally used for reliability of such scales is a stringent one: there should be at least an 80% match between the ratings of the subjects and the analysts; that is, the average number of agreements of the analysts judgements with the subjects should exceed 8 out of 10. This criterion was just met in the present experiment, in which the average agreement of the six subjects with the analysts' judgement was .801. This result supports the conclusion that this is indeed a reliable scale for degrees of mitigation. A more detailed analysis of the data (see [Structural Semantics 83]) suggests that the variance among subjects is due to regional dialect differences and to their being less well trained than the analysts.

2.3.6 Topic and Topic Failure

A careful definition of topic is necessary to investigate why crew members sometimes fail to recognize or continue newly proposed topics, often topics of great operational importance. **Topic** is defined as the propositional content of a speech act. The **propositional content** is what a sentence predicates about the world, what the sentence is about, independent of its social force. For example, (10), (11), and (12) have different social force but the same propositional content.

(10) **Close the window.**
(11) **The window is closed.**
(12) **Is the window closed?**

Using this definition, we have been able to determine instances of topic failure in our sample. We count as topic failed any speech acts expressing a new topic not followed by a speech act having the same topic from another speaker. We have also given a taxonomy of the major topics found (see [Structural Semantics 83]).

2.3.7 Draft Orders and Ratification

Plans are a major means by which a crew can discuss possible actions. A crucial question about this process is how decisions about what actions to take are actually made and expressed. This is a complex social process, requiring appropriate communications among the individuals involved, and depending, in part, on the fact that there is a strict social hierarchy, in which all the participants are highly trained and are moreover legally responsible for the correctness of the decisions made.

Studying the execution of plans means understanding planning as part of the command and control system. From the command and control perspective, a plan is a directive whose propositional content contains possible actions. If such a directive is made by someone other than the captain, or by the captain as a suggestion rather than as an order, then it must be **ratified** before it has the social force of an action which the crew understands is to be performed. Since the final authority rests with the captain, all possible non-routine actions should flow through him for ratification. Examination of the transcripts shows that such ratifictions can be either explicit or implicit. Thus, an action proposed by someone other than the captain may be viewed as a **draft order**, which requires the captain's ratification to turn it into an actual order. Actions proposed but not ordered by the captain are more complex; they may receive approval or modification by crew members, and then flow back to the captain for actual ratification. Under this description, all ratified actions are seen as orders issuing from the captain. These concepts are used to formulate two variables used in our hypotheses, whether or not a speech act is a draft order, and whether or not it is subsequently ratified.

This area is interesting because of its relevance to air crew coordination. A general problem here is how it can happen that important and relevant actions are not in fact taken. One specific form of this is that an appropriate action is actually proposed but then not ratified.

2.4 Level of Significance and Statistical Tests Used

This subsection discusses the statistical tests and levels of significance used in the hypothesis testing.

In statistical research in linguistics and sociology, a .05 level of significance is standard [Herdan 66]. The reader who is not familiar with these disciplines should note that verifying hypotheses in these areas is in general more difficult than verifying hypotheses about physical science data, in part because there are many more sources for uncontrolled variation. Although we have adopted this convention, it should be noted that a significance level of .03 would suffice for all hypotheses actually accepted here.

2.4.1 Assumptions Underlying Use of the t Test

Only two statistics are used for testing the hypotheses in this research, Student's t statistic and the χ^2 statistic. Both statistics test whether or not two samples differ significantly in the values of some variable. The choice of statistic for testing a given hypothesis is determined by whether or not certain assumptions are satisfied by the data. Modern statistical practice has found both the t and χ^2 statistics to be remarkably robust, so that only approximate satisfaction of their underlying assumptions is required [Bowen & Weisberg 80]. Whenever appropriate, the t statistic is preferable to the χ^2 statistic, since the t statistic yields a more definitive decision on the same data. This research only uses the one sided t test, but uses both one and two sided χ^2 tests.

According to the classical view (e.g., [Siegel 56]), appropriateness of the t statistic depends upon approximate satisfaction of four conditions:

(1) the dependent variable has a normal distribution for each of the two populations being compared;
(2) these distributions have equal variance;
(3) the two samples being compared are independent; and
(4) the dependent variable has values on an interval scale.

Let us consider each of these assumptions in relation to the data involved in this study, and in the light of more modern views. Assumption (1) is usually valid for reasonably large samples, and in fact is approximately satisfied by the mitigation scores obtained in this study. Regarding assumption (2), we have computed the variances of each sample for all the hypotheses tested in the research reported here, and have noted that they are approximately equal. (This could be tested using the F statistic, but we have not done so.)

The independency assumption (3) is more problematic because our units of analysis are speech acts rather than individuals. For some hypotheses, the speech acts in the samples compared are generated by different individuals, while for others they are generated by the same individuals in different situations. We have therefore used computational formulas for related- or single- sample (i.e., pooled variance) comparisons. (However, the outcomes should be virtually identical to those for independent sample test procedures.)

The role of assumption (4) is very controversial, and many writers do not believe it is necessary [Gaito 80]. We can argue that the mitigation/aggravation levels of speech acts approximate an interval scale, specifically a scale of just noticeable differences of mitigation/aggravation. If this argument is accepted, then assumption (4) is satisfied whenever the dependent variable is mitigation/aggravation score, and therefore the t test can be used for all hypotheses except 4, 5 and 6. To show that the intervals of the scale of mitigation/aggravation are "jnd's," trial studies were run using two scales having more levels of both mitigation and of aggravation, a first with three levels of each, and a second with one level of aggravation and three levels of mitigation; both had a single "direct" level. It was found that reliable coding could not be achieved using these finer scales. This suggests that the four level scale finally found to be reliable (see Section 2.3.5) is a scale of "jnd's of mitigation/aggravation level." If this is the case, then the scale of mitigation/aggravation is an interval scale whose unit is one jnd of mitigation/aggravation. This argument is not entirely conclusive, because the earlier attempts at reliable scaling with more levels were not as rigorous as our final experiment, and we did not try to determine directly whether or not these levels are really jnd's. One can also follow [Gaito 80] and others in claiming that use of the t test does not require satisfaction of the interval scale assumption[5]. The reader who does not accept our arguments may prefer to use the χ^2 test for each hypothesis. These are given in Section 3.2, and support all our results except in the case of Hypothesis 1.

2.4.2 Assumptions Underlying Use of the χ^2 Test

The χ^2 test must be used for Hypotheses 4, 5 and 6 since the dependent variable in these hypotheses takes the two values "planning or explanation" and "not planning or explanation." These two values do not form an interval scale (in fact, they do not even form an ordinal scale, but only a nominal scale, because it makes no sense to ask whether "planning or explanation" is greater than "not planning or explanation"). The only assumption that needs to be satisfied is that samples from the two distributions are independent. This is clear when the independent variable is rank, since the sets of speakers are then disjoint in the two groups being tested for difference; this justifies the use of this test for Hypothesis 4. For Hypotheses 5 and 6, the independent variables are CRE/non-CRE and CRP/non-CRP, respectively. Although we are unable to give a definitive justification for the applicability of the χ^2 test for these hypotheses, we can give an argument that may be reasonably convincing: because of the relative stability of linguistic frequency distributions, the relatively large numbers of speech acts and speakers, and their relative independence of speaker[6], especially for such a close-knit community as commercial air transport crews, it may be expected that the average rate of planning or explanation (which is the dependent variable) over a number of individuals will also be stable.

3. RESULTS

The first subsection below precisely states eight research hypotheses about the use of language in this setting, reports the results of testing them, and discusses their significance for aviation safety. The next subsection summarizes the results, while the final subsection discusses their generalizability.

3.1 Hypotheses and Test Results

This subsection precisely formulates the null hypothesis and dataset involved in each of our eight research hypotheses, and also gives the statistical test used and the level of significance obtained. The choice of hypotheses to be tested was influenced by the pioneering work of [Foushee & Manos 81]. Each hypothesis is restricted to speech acts whose speaker and addressee are both crew members, because we are not studying air-to-ground communication, nor are we studying communication with flight attendants or passengers. They are restricted to operationally relevant speech acts because there is more linguistic variation in the non-operationally relevant portions of the text, and because non-operationally relevant speech acts are less important for our purpose. Checklist speech acts are excluded because checklist activity is highly stereotyped; in particular, these speech acts are almost always direct and almost never acknowledged. These restrictions apply to all eight research hypotheses and are not repeated below. A further requirement is the well-definedness of the variables occurring in a given hypothesis; for example, speech acts with unknown speaker cannot be used in testing hypotheses that involve speaker rank. Frequency tables for each hypothesis are omitted here, but may be found in [Structural Semantics 83]. Some further hypotheses, which we were unable to test on the present sample, may be found in [Structural Semantics 83].

---

[5] Gaito cites Lord as saying "The numbers do not know where they come from."

[6] This argument is not circular, because the tests supporting the homogeneity of the sample (see Section 3.3) use the t test.

3.1.1 Requests to Superiors Are More Mitigated

The null hypothesis here is that the mean mitigation/aggravation score for requests to subordinates equals (or exceeds) the mean score for requests to superiors.

This test is limited to requests because requests (which include orders, questions, draft orders and suggestions) are the most characteristic speech act in command and control discourse, and also because the consequences of misunderstanding requests are more direct and immediate than those of any other speech act. This hypothesis represents the intuition that the speech of subordinates is more tentative and indirect than the speech of superiors. The hypothesis is important because it posits a direct effect of the basic social hierarchy on cockpit discourse. If this hypothesis is verified, and if it is also shown that more highly mitigated speech acts are more often misunderstood or ignored (as is strongly suggested by the acceptance of Hypotheses 7 and 8 below), then it should be worth testing whether training subordinates to use less mitigation would improve crew performance. Such a training hypothesis can not itself be tested with data from accident transcripts, but could be tested with simulator experiment data.

Because the hypothesis asserts that one mean is greater than another, it is tested with a one sided Student's t test. The frequency data for this hypothesis from the six test group transcripts yields t=2.38 (df=136 and p=.009), using the normal approximation, which is valid because of the large sample size. The hypothesis is therefore accepted, and we conclude that crew members indeed use more mitigation in making requests to superiors in the test transcript sample. Testing the hypothesis with speech acts from the two hypothesis formulation transcripts yields a similar pattern of frequencies, but with an obtained probability of only .32. The hypothesis is therefore not supported by these data, perhaps because there are too few speech acts to achieve the desired significance level. However, because the hypothesis has been accepted on data from the test transcripts, the speech acts from the two groups can be combined. The pooled frequencies yield t=2.01 (df=252, p=.022), so the hypothesis is accepted for the entire dataset.

Since appropriateness of the parametric t test depends on homogeneity of variance, it is interesting to notice that in this dataset, the two samples involved do indeed have approximately equal standard deviations. For speech acts from the six transcripts in the test group, the standard deviation of speech acts by subordinates is .516, while that of speech acts by superiors is .579.

3.1.2 Requests Are Less Mitigated in Crew Recognized Emergencies

The null hypothesis here is that the mean mitigation/aggravation score for requests in CRE equals (or exceeds) the mean mitigation/aggravation score for requests not in CRE.

This hypothesis reflects the intuition that when crew members know that they face an emergency situation, their speech is less tentative and indirect. It is based on the notion that in any utterance, the speaker is encoding both his understanding of the situation he is talking about (the propositional content) and his understanding of the relation between himself and his addressee. Mitigation level is a major linguistic means by which a speaker can indicate his understanding of this social relation. When the situation becomes urgent, we might expect the speaker to focus most of his attention on it, and thus less attention upon social relations.

Verification of this hypothesis would imply that crew members are able to vary their level of mitigation depending on their perception of the circumstances. This should mean that training crew members to use less mitigation in specified circumstances would not seem new or strange to them, because mitigation level is already something that they alter when aware that they are in an emergency situation. Under the assumption that what experienced crews do in emergency situations may be valuable, verification of this hypothesis would also lend some support to the hypothesis that training crews to speak more directly would improve their performance and thus reduce accidents (however, caution is advisable in drawing such a conclusion from the present dataset of accident transcripts).

The frequencies obtained from the test transcripts for investigating this hypothesis yield t=3.05 (df=166, p=.001), and the hypothesis is therefore accepted. The obtained probability level for similar comparisons of speech acts in the hypothesis formulation group of transcripts is .026. It is therefore permissible to combine the two datasets, yielding t=3.46 (df=276, p=.0003). Hypothesis 2 is therefore very strongly supported.

3.1.3 Requests are Less Mitigated in Crew Recognized Problems

The null hypothesis here is that the mean mitigation/aggravation score for requests in CRP equals (or exceeds) to the mean mitigation/aggravation score for requests not in CRP.

This hypothesis corresponds to the intuition that crew members' speech is less tentative and indirect when they know they face a problem. Its significance is similar to that of the previous hypothesis. (Note that every CRE speech act is also a CRP speech act.)

The frequencies obtained from speech acts in the test group of transcripts, comparing CRP and non-CRP mitigations levels, give $t=2.34$ ($df=166$, $p=.010$). The hypothesis is therefore accepted for the test dataset. For the hypothesis formulation transcripts, the corresponding obtained probability level is .149. Combining the two groups produces $t=1.79$ ($df=276$, $p=.047$). The research hypothesis is therefore accepted for the dataset as a whole.

3.1.4 Captains and Subordinates Differ in Frequency of Planning and Explanation

The null hypothesis is that the percentage of speech acts in explanation and planning discourse units produced by subordinates equals the percentage produced by superiors.

This research hypothesis indirectly probes the effects of social hierarchy on subordinates' contributions to explaining what is happening and to planning what should happen in the future. Accepting the research hypothesis would suggest that the social hierarchy might be having a detrimental effect on crew communications.

We use a two sided χ^2 test with one degree of freedom. Discourse type frequencies for speech acts in the six test transcripts yield $\chi^2=1.52$ for an obtained probability level of .22, not suporting the research hypothesis. A similar evaluation of speech acts from the formulation group transcripts gives $\chi^2=1.13$, with probability level .29. It is therefore permissible to combine the two datasets, yielding $\chi^2=2.97$ with $p=.086$. Thus the null hypothesis cannot be rejected on the pooled data, and therefore the research hypothesis is not accepted.

Modern management theory generally asserts that groups are more effective when subordinates contribute more than superiors. Moreover, informal examinations of accident transcripts have suggested to many observers that captains can behave in an autocratic manner that prevents subordinates from making appropriate contributions. Our results suggest that it would be valuable to determine whether crew performance is improved by training subordinates to engage in more planning and explanation, and training captains to encourage this, at least in the condition of CRP but not CRE. In this connection, it would be important to determine if there are circumstances, such as CRE, in which it would be counterproductive to engage in more planning and explanation.

If we had been *a priori* certain of the direction of difference in frequency of planning and explaining between captains and subordinate crew members, we could have used a one sided χ^2 test, and the hypothesis that captains plan and explain more would have been accepted with an obtained probability level of .043. Subordinates in fact only produced 38% of the planning and explaination speech acts in the pooled dataset, while captains produced 62%; in fact captains and subordinates each produced about half of all speech acts in this dataset, but planning and explanation speech acts are only 9% of the total. It would also have been interesting to test whether captains plan and explain more during CRP and during non-CRE, but we have not done so.

3.1.5 Planning and Explanation Are Less Common in Crew Recognized Emergencies

The null hypothesis is that the percentage of speech acts that occur in planning and reasoning discourse units in CRE equals (or exceeds) the percentage that occur in non-CRE.

This hypothesis represents the intuition that when crew members are aware that they face an emergency, they do less planning and explaining, because an emergency calls for immediate action. Precise knowledge of the distribution of planning and explanation in accident transcripts is important because it may suggest circumstances in which crews should be trained to do more planning and explanation, or else less, when it proves to be counterproductive.

Because the research hypothesis asserts the degree of difference and there is only one degree of freedom, a one sided χ^2 test is used. The speech act frequencies for this hypothesis in the test transcripts yield χ^2=3.87 for an obtained probability level of .025. The hypothesis is therefore accepted at the .05 significance level. The corresponding test for speech acts from the hypothesis formulation transcripts yields χ^2=7.03 (p=.004). Thus, it is permissible to combine the two datasets for Hypothesis 5. The combined frequencies yield χ^2=12.49 (p=.0002); the hypothesis is therefore strongly supported on the pooled data. Further discussion of the implications of this result is included with that of the following hypothesis

3.1.6 Planning and Explanation Are More Common in Crew Recognized Problems

The null hypothesis is that the percentage of speech acts that occur in planning and reasoning discourse units in non-CRP equals (or exceeds) the percentage in CRP.

This hypothesis corresponds to the intuition that crew members use more planning and explanation when they are aware that they face a problem. If verified, this hypothesis would strengthen our confidence in the relevance of the variables involved (discourse type and CRP), and would also confirm the value of training crews to plan and reason in problem situations.

Again using a one sided χ^2 test (df=1), the discourse type frequencies obtained from speech acts in the test transcripts yield a $\chi^2=25.90$, with an obtained probability level beyond .000001. The hypothesis is therefore very strongly confirmed in this dataset. The corresponding χ^2 value for discourse type frequencies from the hypothesis formulation transcripts is .27, for an obtained probability level of .30. Frequencies by discourse type for speech acts from the combined group of eight transcripts yield $\chi^2=12.03$, and an associated probability level of .0003. The hypothesis is therefore strongly confirmed for the entire dataset.

This result taken together with the findings relevant to Hypothesis 5 suggests that, perhaps contrary to expectation, more planning and reasoning occur when the crew believes that it is dealing with a problem, but not when it believes that it is dealing with an emergency. One explanation for this result is that by the time an emergency situation has developed, crew members may feel that it is too late to take the time to plan as a group, or to explain the reasons for taking specific actions. It is of course possible that more planning and explanation would be desirable in some emergency situations, but not in others. This suggests using simulator experiments to determine in which flight segments (if any) more planning and explanation produce better performance. In any case, these results make it clear that crews should plan as effectively as possible during CRP, because they may not have time for planning during a subsequent emergency.

It should be noted that because this study is based upon accident transcripts, it cannot be assumed that observed crew behavior in this data is necessarily optimal. It seems quite possible that the data used in this study are a combination of good and bad instances of cockpit planning and reasoning; thus, testing the present hypothesis on data from normal flights should yield more definitive results.

3.1.7 Topic Failed Speech Acts Are More Mitigated

The null hyothesis is that the mean mitigation/aggravation score for speech acts whose topic has failed is greater than or equal to that for speech acts whose topic has succeeded.

This hypothesis and the next one attempt to probe the idea that excessive mitigation can have undesirable effects in the cockpit. Since the effect of mitigation on performance (e.g., the probability of an accident) cannot be explored directly with the present data, we are forced to examine less direct connections. This hypothesis represents the intuition that a new topic is less likely to be continued by its addressees if the speech act in which it is introduced is excessively mitigated.

A comparison of mitigation scores for the two topic conditions using speech acts from the six test transcripts gives t=1.65 (df=182, p=.01), and thus this hypothesis is accepted. For comparisons based on the hypothesis formulation transcripts, t=2.23 (df=80, p=.013). Examining the combined dataset mitigation levels across topic conditions in all eight transcripts yields t=2.493 (df=264, p=.0064). Therefore the hypothesis is accepted.

This result lends strong support to the intuition that excessive mitigation can have undesirable effects on crew performance. A number of NTSB reports have recommended assertiveness training for crew members to encourage more effective participation by subordinates. (See, for example, [NTSB 79].) Verification of the present hypothesis and the following one demonstrate effects for one kind of lack of assertiveness. Moreover, this kind of lack of assertiveness is defined precisely enough to allow both for training and for the evaluation of training methods.

3.1.8 Unratified Draft Orders Are More Mitigated

The null hypothesis is that the mean mitigation/aggravation score for ratified draft orders equals (or exceeds) the mean for draft orders that are not ratified.

This hypothesis attempts to test the intuition that when a crew member proposes a suggestion to the captain, the more indirect and tentative that suggestion is, the less likely the captain is to ratify it. Statistical evaluation of the frequencies for ratified and unratified draft orders from the six test transcripts yields t=2.927 (df=45, p=.002). The hypothesis is therefore accepted for speech acts from the test transcripts. For similarly classified speech acts from the hypothesis formulation transcripts, t=.589 (df=13). The t statistic table gives an obtained probability level of approximately .2, so the two groups can be combined, yielding t=2.412 (df=60, p=.008) on the pooled data. Thus, this hypothesis is strongly supported.

Like Hypothesis 7, this hypothesis implies that excessive mitigation can have undesirable effects on crew performance. In particular, this hypothesis focusses attention on the situation in which a subordinate makes a correct suggestion which is ignored. Training in linguistic directness may be valuable in correcting this kind of pattern.

3.2 Summary of Results

This subsection gives two figures showing first, the independent and dependent variables that are used in each hypothesis, and second, the results of testing each hypothesis.

	independent variables				
dep vbls	rank	CRE	CRP	topic failed	ratif
mitigatn	1	2	3	7	8
plan/expln	4	5	6		

Figure 7: Hypotheses with Dependent and Independent Variables

Hypothesis	N	t	χ^2	df	P_t	P_χ	Decision
1	254	2.01	7.45	3	.022	.05+	Yes
2	278	3.46	12.81	3	.0003	<.01	Yes
3	278	1.79	4.70	3	.047	<.01	Yes
4	879		2.97	1		.086	No
5	1039		12.49	1		.0002	Yes
6	1039		12.03	1		.0003	Yes
7	266	2.49	7.95	3	.0064	<.05	Yes
8	62	2.41	9.52	3	.008	.02+	Yes

Figure 8: Summary of Results

Figure 7 shows the independent and dependent variables occurring, and which hypothesis uses each. (The two blanks suggest possibly interesting hypotheses that have not been tested in this study.) Figure 8 shows for each hypothesis: the size, N, of the dataset used to test it (in each case this includes speech acts from all 8 transcripts); the obtained t value (if any); the obtained χ^2 value; the number of degrees of freedom (for the χ^2 test); the obtained probability level for the t test; the obtained probability level for the χ^2 test; and the decision (whether or not the research hypothesis was accepted). The decisions obtained using the χ^2 test agree with those obtained using the t test, except in the case of Hypothesis 1. Although the χ^2 value is very close to that required for acceptance, a reader who is doubtful about the applicability of the t test, may want to consider this hypothesis rejected.

3.3 Representativeness of the Sample

We now consider the **generalizability** of our results from these transcripts to the broader population of commercial aviation cockpit discourse. The results will generalize provided that the sample is representative. This subsection presents various arguments for the representativeness of our sample.

It might be argued that the sample cannot be representative because it consists of only eight transcripts. But this is a misunderstanding of the nature of the sample. It consists not of the eight transcripts (or equivalently, the eight crews), nor of the 25 speakers present in the transcripts, but rather of all the (operationally relevant) speech acts produced by these speakers. This is a much larger sample, and one which is much more likely to be representative of its population, for reasons given below.

The first and most basic argument for representativeness is that a sample is very likely to be representative if it is sufficiently large and is also a random sample; in fact, the probability that a random sample is not representative can be made as small as desired by making the sample large enough. The argument that our sample is large enough is based upon experience with statistical studies of other linguistic data. This experience suggests that samples of size one or two hundred units are generally adequate [Herdan 66] and that smaller samples often suffice if the pattern of variation is not especially complex [Guy 80]. Only Hypothesis 8 might be in doubt on the ground of sample size; however, as this hypothesis does not seem especially subtle, there seems to be no cause for more than raising a mild cautionary flag.

We now give three arguments for the randomness of our sample. The first and most direct argument is that our sample is random because the criteria used for transcript selection are in fact *statistically independent* of the dependent measures used in the hypotheses. For example, it seems clear that whether or not a critical segment occurs in a transcript cannot effect the mitigation level of the speech acts occurring in that transcript. (Section 2.1.1 gives the selection criteria used in this study.)

The second argument for randomness of the sample is based on the *principle of locality of effects*, which says that, although there are significant sequential dependencies in natural language, they are largely confined to units a few steps earlier in the sequence, and hence have little effect on the randomness of any reasonably large sample. This has been observed for units at all levels of the linguistic hierarchy, including phonemes, morphemes, lexemes (i.e., words) and syntactic phrases, and it is presumed to hold for speech acts as well, although this has not been formally tested.

A third argument for representativeness is the fact that the sample can be successfully used as a standard of comparison for the behavior of crew members, that is, *the sample is homogenous*. Since we have aggregated data from a number of speakers, it might be questioned whether the sample is dominated by a few loquacious speakers who exhibit unusual linguistic behavior. To support the assertion that individual differences are relatively unimportant in this sample compared to systematic differences arising out of the cockpit situation in which the language is produced, we have tested whether or not the most loquacious speaker of each rank differs significantly from his colleagues of the same rank, using the most important, and perhaps the most sensitive, measure in the research reported here, namely degree of mitigation/aggravation. We have found that, for speakers of a given rank, the sample is not dominated by a few speakers with unusual linguistic behavior. (For more details, see [Structural Semantics 83].)

As a final point, recall from Section 2.1.3 that the transcripts are divided into two disjoint groups, called the hypothesis formulation and test groups, in order to reduce the likelihood that the result obtained from testing a given hypothesis is due to some uncontrolled variable, different from the independent variable of the hypothesis in question.

The results of the statistical tests on the research hypotheses of this study are clearly valid as *descriptive* statistics, that is, as statistical summaries of a particular sample. The above arguments for generalizability support our giving these results the usual *inferential* interpretation.

4. CONCLUSIONS AND EXTENSIONS

The results reported above support the conclusion that a methodology is now available for the detailed analysis of cockpit discourse, and that this methodology can be applied to improving aviation safety. This methodology has produced a number of verified results concerning the linguistic behavior of air crews that seem to have significant implications for crew training. In addition, a number of interesting directions for further research have been suggested. The first two subsections below detail what we think have been the main contributions of this research, while the third discusses its extensions.

4.1 General and Basic Contributions

1. A theory of the structure of command and control discourse that includes a determination of its relationships to planning and explanation, as well as determination of its basic speech acts, which are request, report, acknowledgement and declaration.
2. A general theory of the structure of discourse; this theory involves analyzing a given discourse unit as a sequence of transformations that construct an underlying tree structure representing the structure of the discourse, i.e., a hierarchical classification of the discourse parts and their relationships.
3. A classification of the discourse types that occur in aviation discourse. These are: command and control chain, including the subtype of checklist; planning; explanation; and narrative and pseudo-narrative.
4. A scale of **mitigation** levels for speech acts occurring in aviation discourse. This five point scale ranges from "highly mitigated" to "aggravated" and has "direct" as its zero point. The scale has been experimentally validated.
5. A theory of draft orders (suggestions for action that have not yet been ratified by the captain) and how they come to be ratified has been developed, based on the theories of planning, explanation, and command and control discourse.
6. A collection of variables has been isolated that summarize many important characteristics of the speech acts that occur in cockpit discourse.

4.2 Applied and Specific Contributions

This subsection describes what we believe are the relation of the hypotheses discussed here to further research and direct training in crew coordination and communiation for aviation safety. It should be remembered that these results are necessarily limited by the restriction of our data to accident transcripts. It should be possible to go much further in the directions indicated here when both systems data and non-accident data are available.

1. It has been shown that the average mitigation level of requests by subordinates is significantly higher than that of requests by superiors. It would be important to test whether this asymmetry contributes to the misinterpretation of suggestions and commands in the cockpit, because it should not be difficult to train subordinate crew members to use less mitigated language, or (as NTSB reports put it) to be more assertive.
2. It has been shown that requests are less mitigated during a Crew Recognized Problem, and are still less mitigated during a Crew Recognized Emergency. This suggests that crew members should not find it strange or abnormal to be trained to use less mitigation, since variation of mitigation level is something that they already do under certain conditions. It also suggests that training in linguistic assertiveness would only be reinforcing a tendency that already appears under problem and emergency conditions.
3. It has been shown that superiors produce a higher proportion of explanation or

planning speech acts than subordinates. The optimal ratio is not clear, and may depend on the context; it would be important to investigate this. There are reasons to believe that this ratio would be a good indicator of degree of the authority delegated by a given captain to his crew.

4. It has been shown that planning and explanation are much more common during crew recognized problems, and that they are *less* common during crew recognized emergencies. This suggests further research to discover whether training crew members to engage in more planning and reasoning under real emergency conditions would improve performance.
5. It has been shown that more mitigated speech acts introducing a new topic, are less likely to have their topic become the subject of further conversation. This demonstrates the importance of crew members not using mitigated language when introducing operationally significant topics. Because this also is presumably behavior for which crew members can be trained, it would be interesting to explore both the basic linguistic phenomena further, and to test whether or not such training can improve any objective performance measures.
6. It has been shown, with a very high level of significance, that on the average, draft orders that do not get ratified are more mitigated than those that do get ratified. The implications of this result are very similar to those of the previous result, but concern the ratification of subordinates' suggestions rather than the success of their topics.
7. The research reported here (see [Structural Semantics 83] for details) suggests that a number of other linguistic variables should be investigated for correlation with objective system and crew performance variables. These variables include: degree of command and control coherence, as defined in; the rate of request-report-acknowledge triples; the rate of planning and reasoning; and the rate of simple acknowledgements. In certain cases, it might be less costly to use a reliable linguistic variable as an indicator of some objective performance measure than to measure it directly. In other cases, important training implications might be discovered.
8. Finally, the research program initiated in this report should have many applications to the design of aviation procedures and equipment systems that involve communication (such as onboard display and speech generation). This possibility of application arises from the clear demonstration that air crew discourse involves definite linguistic structures, and that these structures correspond in specific ways to the operational structure of the flight. This means that there are only certain times when it is natural for certain kinds of communications to occur, and that there are natural forms for each kind of communication. For example, a piece of equipment in the cockpit that produced complex verbal information about the status of the flight plan would probably not be useful unless it produced this information at the right time and in the right form.

4.3 Extensions

The methodology described here is presently being used in a study of crew coordination factors and their relationship to flight task performance, using as data audio and video recordings of 16 full mission simulations [Murphy *et al.* 84]. This will permit some important extensions of the present research including:

1. Extension of the hypotheses to non-verbal performance and to factors of the aircraft systems.
2. Comparison of communication in successful and unsuccessful flight performance.

We expect that the findings of the current study will be confirmed, refined and extended by this richer dataset. It is hoped that this will lead to the development of new methods for training crews in more effective communication, and will provide guidelines for the design of improved aviation procedures and equipment.

References

[Bowen & Weisberg 80]
Bowen, B. D. and Weisberg, H. F.
An Introduction to Data Analysis.
Freeman, 1980.

[Brown and Levinson 79]
Brown, Penelope and Levinson,Stephen.
Universals in Language: Politeness Phenomena.
In Esther N. Goody (editor), *Questions and Politeness: Strategies in Social Interaction*, . Cambridge University Press, 1979.

[Foushee & Manos 81]
Foushee, H. Clayton and Manos, Karen L.
Information Transfer within the Cockpit: Problems in Intracockpit Communications.
Technical Report, NASA Ames Research Center, 1981.
In NASA Technical Paper 1895, *Information Transfer Problems in the Aviation System*, edited by C. E. Billings and E. S. Cheaney.

[Gaito 80] Gaito, John.
Measurement Scales and Statistics: Resurgence of an Old Misconception.
Psychological Bulletin 87(3):564-567, 1980.

[Gazdar 79] Gazdar, Gerald.
Pragmatics: Implicature, Presuppostion and Logical Form.
Academic Press, 1979.

[Goguen, Weiner & Linde 83]
Goguen, Joseph, Weiner, James and Linde, Charlotte.
Reasoning and Natural Explanation.
International Journal of Man-Machine Studies 19:521-559, 1983.

[Gordon & Lakoff 71]
Gordon, David and Lakoff, George.
Conversational Postulates.
Papers from the Regional Meeting, Chicago Linguistics Society , 1971.

[Guy 80] Guy, Gregory.
Variation in the Group and the Individual: The Case of Final Stop Deletion.
In William Labov (editor), *Locating Language in Time and Space*, . Academic Press, 1980.

[Herdan 66] Herdan, Gustav.
The Advanced Theory of Langauge as Choice and Chance.
Springer-Verlag, 1966.

[Labov 72] Labov, William.
The Transformation of Experience into Narrative Syntax.
In *Language in the Inner City*, . University of Pennsylvania Press, Philadelphia, 1972.

[Labov & Fanshel 77]
Labov, William and Fanshel, David.
Therapeutic Discourse; Psychotherapy as Conversation.
Academic Press, 1977.

[Linde 74] Linde, Charlotte.
The Linguistic Encoding of Spatial Information.
PhD thesis, Columbia University, 1974.

[Linde & Goguen 78]
Linde, Charlotte and Goguen, Joseph.
The Structure of Planning Discourse.
Journal of Social and Biological Structures 1:219-251, 1978.

[Linde & Labov 75]
Linde, Charlotte and Labov, William.
Spatial Networks as a Site for the Study of Language and Thought.
Language 51, 1975.

[Murphy 80] Murphy, Miles.
Analysis of Eighty-four Commercial Aviation Incidents: Implications for a Resource Management Approach to Crew Training.
In *Proceedings, Annual Reliability and Maintainability Symposium*, . IEEE, 1980.

[Murphy *et al.* 84]
Murphy, M., Randle, R., Tanner, T., Frankel, R., Goguen, J., Linde, C.
A Full Mission Simulator Study of Aircrew Performance: The Measurement of Crew Coordination Factors and Their Relation to Flight Task Performance.
In Hartzell, E. James, and Hart, Sandra (editors), *Papers from the 20th Annual Conference on Manual Control*, . NASA Ames Research Center, 1984.

[NTSB 79] National Transportation Safety Board.
Aircraft Accident Report -- United Airlines, Inc., McDonnell-Douglas DC-8-61, N8082U, Portland.
Technical Report, National Transportation Safety Board, 1979.

[Sachs 74] Sachs, Harvey.
An Analysis of the Course of a Joke's Telling in Conversation.
In Bauman, R. and Sherzer, J. (editor), *Explorations in the Ethnography of Speaking*, . Cambridge University Press, 1974.

[Searle 79] Searle, John.
Expression and Meaning.
Cambridge University Press, 1979.

[Siegel 56] Siegel, S.
Nonparametric Statistics for the Behavioral Sciences.
McGraw-Hill, 1956.

[Structural Semantics 83]
Goguen, Joseph and Linde, Charlotte.
Linguistic Methodology for the Analysis of Aviation Accidents.
Technical Report, Structural Semantics, 1983.
NASA Contractor Report 3741, to Ames Research Center.

N85
14556

UNCLAS

D21

N85 14556

A FULL MISSION SIMULATOR STUDY OF AIRCREW PERFORMANCE: The Measurement of Crew Coordination and Decisionmaking Factors and Their Relationships to Flight Task Performance

Miles R. Murphy, Robert J. Randle, Trieve A. Tanner, NASA/ARC
Richard M. Frankel, Wayne State University
Joseph A. Goguen and Charlotte Linde, Structural Semantics

Abstract: Sixteen three-man crews flew a full-mission scenario in an airline flight simulator. The scenario was designed to elicit a high level of verbal interaction during instances of critical decisionmaking. Each crew flew the scenario only once, without prior knowledge of the scenario problem. Following a simulator run and in accord with formal instructions, each of the three crewmembers independently viewed and commented on a videotape of their performance. Two check-pilot observers rated pilot performance across all crews and, following each run, also commented on the video tape of that crew's performance. A linguistic analysis of voice transcripts is being made to provide added assessment of crew coordination and decisionmaking qualities. Measures of crew coordination and decisionmaking factors are being correlated with flight task performance measures. Some results and conclusions from observational data are presented.

INTRODUCTION

Crew coordination and decisionmaking have been implicated as factors in many accidents and incidents (NTSB, 1976; Murphy, 1980) and many crew factors have been suggested as causes of ineffective crew performance, for example, pilot-copilot role relations, lack of decisive command, and strained social relations (Murphy, 1977). A full-mission simulator study of crew performance (Ruffell-Smith, 1979) related the ineffective management of both human and material resources to increased decision times. Generally, however, suggested causal factors in air crew performance effectiveness have not been well defined through systematic study or research. One reason for this status is suggested to be a lack of an effective method for isolating and quantifying crew interaction factors and for relating these factors to flight task performance (cf. Murphy, 1977; Hackman and Morris, 1975). This study continues efforts to develop such methods (Goguen et al, 1984; Goguen and Linde, 1983; Foushee and Manos, 1981; Murphy, 1980).

The primary objective of this study is to develop methods for quantifying crew coordination and decisionmaking factors and their relationships to flight task performance. A secondary objective is to develop information about crew process and performance for application in the development of resources management training programs. Of special interest is information on how errors evolve in the cockpit, particularly errors involving interpersonal factors.

Relationships between several crew and systems performance measures and some personal and crew process variables are explored in this study. Personal variable categories include personality and background variables such as age and experience, for example. The primary emphasis , however, is on crew process, or interpersonal interaction. Constructs or variable classes of major concern here are suggested, a priori, to be: 1. command hierarchy, 2. command style, 3. interpersonal communications, 4. crew coordination, 5. resources management, and 6. group decisionmaking.

The overall design of this study is similar to that of NASA's first full mission simulator study of air crew performance (Ruffell-Smith, 1979). This study differs from the earlier study mainly in the more formal and comprehensive assessment of interpersonal interaction factors, in scenario design, and in crew member selection criteria.

METHOD

Simulator

A Boeing 720B flight training simulator, an early version of the Boeing 707, was used in the study. This simulator is an FAA approved visual simulator with a model-board scene and has three degrees of freedom in motion: pitch, roll, and heave. The simulator is operated by the Airline Training Institute (ATI), San Carlos, California.

Air Traffic Control

A current, professional air traffic controller was used in the simulation. The controller also participated with another member of the experimental team in simulating conversations with other aircraft, thus providing background conversations on the ATC network.

Scenario

Simply, the scenario represented a flight from Tucson (TUS) to Los Angeles (LAX) via Phoenix (PHX) with a forced diversion to an alternate upon reaching LAX. The crew's enactment of the scenario began with a Captain's Briefing in the simulated operations room at TUS and ended upon deplaning at the selected alternate---either Palmdale (PMD) or Ontario (ONT).

The scenario was designed to evoke a series of decisions about where to proceed following a missed approach at LAX due to a nose gear not-down-and-locked indication. This situation was exacerbated by having occurred at a time when the Los Angeles basin, including Ontario, was experiencing low and deteriorating ceilings and visibilities due to coastal fog. Ontario, located inland from Los Angeles, was lagging Los Angeles in this deterioration. And, just over a mountain range, out of the basin, Palmdale was experiencing clear weather with good visibility. Upon going through a complete gear check procedure taking several minutes, the crews would discover that the gear was down and pinned and they could therefore assume that the panel light indication was faulty.

Within this scenario the most critical dimensions of the decision process involve: 1. when to proceed from the LAX area to an alternate, and 2. the choice of the alternate. Related subsidiary choices involve: 1. whether to do a complete gear check in the LAX area, 2. whether to make a second landing attempt at LAX (ceilings and runway visual range [RVR] degrade to legal minimums at LAX during this choice "window" and will go below minimums if and when the aircraft crosses the outer marker), 3. whether to raise the gear for fuel conservation while flying to the alternate, and 4. whether to declare an emergency for either the gear problem or a minimum fuel problem.

Research Approach

An observational-correlational approach was used in the study. Each crew flew the scenario only once, without prior knowledge of the scenario problem. The intent of the approach and procedure was to maximize a valid description of natural crew performance.

Relationships between several crew and systems performance measures and variables in the following personal and operational categories are assessed.

Personal characteristics: Several personality dimensions were assessed. Instrumentality and expressiveness (e.g., Spence and Helmreich, 1978) were measured to assess their potential as predictors of differences in interaction and performance. These dimensions have been suggested as possible predictors of managerial effectiveness in such concepts as the managerial grid (Blake and Mouton, 1978). A multi-dimensional measure of achievement motivation was also included. These dimensions included attitudes toward work, mastery, and competitiveness (e.g. Spence and Helmreich, 1978).

Background variables: Background data obtained from crewmembers included, for example: length of time in crew position for the B707, time since last proficiency check in the B707, a rating of how often each crewmember had flown with other members of the crew in the past.

Pre-simulation activity variables: Pre-simulation data obtained from crewmembers included, for example, departure date and duration of last duty trip, the quantity and quality of sleep the night before the simulator run.

Crew process variables: This category of variables include those usually classified under the general term "crew coordination"---for example, indicators of concepts such as command structure and command style---as well as the strengths of specific practices such as: informing other crewmembers of intended actions, performing crew briefings early in relation to critical flight events, setting appropriate task priorities, delegating tasks appropriately, etc. This category of variables also includes some components of the crew decisionmaking process---for example, those interpersonal components by which a crewmember suggestion may be "ratified" by the Captain and become a command (cf. Goguen et al, 1984; Goguen and Linde, 1983). Another such example would concern components by which suggested topics are continued or ignored. Data for the assessment of crew process variables are obtained from observer ratings, a linguistic analysis of audio records, and a video peer review process.

Crew and system performance measures, that is, the flight task measures, are derived from observational data and simulator state data. Crew performance measures include, for example: reaction time in detecting a system failure, time to a decision, system operation errors. System performance measures include, for example: flight path control accuracy, altitude excursions on climb-out, excessive flight velocities, etc. There is some evidence that simulator system measures are correlated with measures obtained in the aircraft (Randle, et al, 1981).

Another part of this study, reported separately (Calfee et al, 1984) focuses on how crewmembers represent the task of flying and perceive individual responsibilities. Interview data for these analyses were obtained from crewmembers before and after the simulator run.

Data Acquisition

Simulator state data: Twenty six channels of simulator state data were acquired and recorded using an Analog Devices MACSYM 150 digital measurement and control microprocessor with floppy disk storage. Table 1 describes these discrete and analog signals.

Operational voice recordings: An 8-track audio recorder was used in recording the voices from the radio and intercomm nets, individual crewmember voices from lapel mikes, and a channel containing a composite of the voices.

Video recordings: Four video cameras were located in the simulator, out of sight of the crewmembers. Four individual views were recorded: a frontal, upper torso view of each of the three crewmembers and a context view shot from the back of the simulator cockpit showing all three crewmembers performing at their instrument panels and controls. A fifth kind of recording portrayed these four views as a quad image composite for presentation on a single screen. This composite presentation shows the Captain in the upper left quadrant, the First Officer in the upper right quadrant, and the Flight Engineer in the lower right quadrant; the lower left quadrant portrays the context view, preserving the same relative locations of the crewmembers (see figure 1). Four of these quad image, composite views were recorded real-time for use in the video peer review process.

Video peer review comments: Comments made by each crewmember and observer during their review of the video tape of that crew's performance were recorded on one channel of 2-channel audio cassettes. Greenwich Mean Time (GMT), representing the operations time for the flight, was recorded on the second channel. This enabled the computation of time distribution for all interpretive comments relating to any single event.

IRIG timing system: An IRIG timing system generated the GMT for time-labelling events on all audio and video records. The microprocessor used in acquiring simulator state data was manually synchronized to this GMT.

Observer comments: The two check pilot observers rated performances of the two pilot crewmembers from their location in the rear of the simulator. They rated performance in several categories during each phase of the flight and for each of two flight legs. They rated performance on 5-point scales for which a "3" indicated average performance and a "1" and "5" indicated below and above average performance, respectively. Rated categories included, for example: Crew Coordination/Communications, Planning and Situation Awareness, and Overall Performance and Execution.

Subjects

Selection: The crewmembers were paid volunteers. Their experience represented a wide range in reference to airline of origin and recency, or currency, on B-707 line operations. Some were current on the B-707. Many had recent B-707 line experience and were now flying other jet aircraft in line operations. Some were retired from the line. This diversity in experience was important as an aid in evaluating the sensitivity of the various performance measures. Thus crew composition ranged from one in which all members were retired from the line to one currently flying the B-707 as an intact crew.

Differences training: All crewmembers received six hours of classroom differences training and four to eight hours of simulator differences training. The number of hours of simulator differences training that a crewmember received was based on recency. Subjects were formed into crews prior to this simulator training and were instructed in coordinated procedures during this training.

Baseline system performance data: Some baseline system performance data were obtained for each of the two pilot crewmembers at the completion of the simulator training. When the instructor judged that a pilot had sufficient training, the pilot-subject made three take-offs and landings in the simulated LAX traffic pattern. They included the usual takeoff, climb, vectors to ILS, approach, landing, and roll-out. No further instruction was given during this exercise. All the variables listed in table 1 were continuously recorded during these baseline data runs.

Procedure

Initial briefing: In an initial briefing the potential crewmembers were informed about the study objectives, data-taking procedures, the simulator to be used, and the differences training procedures. They were also told that they would be flying a full mission scenario and that: "The scenario (flight) is initiated at Tucson (TUS) with a Captain's Briefing and is to continue to Los Angeles (LAX) via Phoenix (PHX). The flight will have arrived at TUS from Dallas, the dispatch location, on the day of January 31, 1984 and is scheduled to depart that afternoon. January 31 is specified so that charts, NOTAMS, dawn and dusk times, etc. will be standard across crews. Weather and other factors will be programmed for a hypothetical day in January."

They were also informed about our intentions for participant anonymity. And they were told that: "In order to insure that all crewmembers participating in the study receive identical and similarly timed information, we ask that information about the study be kept confidential until the study is completed. You will be mailed a notice of completion." A copy of the briefing was given to continuing participants, about 97% of this group.

Pre-run instructions and briefing: On the morning of a crew's simulator run they received final instructions on the project procedures---a set of procedures in use at the Airline Training Institute. These final instructions concerned flight plan and dispatch formats and a review of simulator limitations.

The crew then received a final briefing in which they were told: "We would like you to perform just as you would on a regular airline flight, ignoring limitations of the simulator to the extent possible. We ask you not to even use the word 'simulator' while in your role as a crewmember. We ask you to please play your role fully, i.e., do just what you would do in an actual aircraft, even if you sometimes have to 'fake it' because, as you know, full aircraft duplication does not exist." This lack of full duplication mostly involved minor communication and navigation discrepancies on which the crew had been instructed. The one major exception to full duplication was access to a lower bay of the aircraft for inspection of the nose gear pin. For this task simulation of time to wait and results to report, e.g., "pin in", was effected by an observer handing the Flight Engineer an instruction card as he began his simulated enactment of the task.

During this briefing the crew was also informed that operational time (GMT) was to be initiated on this aircraft at the time the Flight Engineer was released from the Captain's Briefing to start his pre-flight checks, and that this time was to be duplicated in the simulated operations room. This time was initiated as one-half hour before block departure time. The crew was also instructed to remain in role on the ground at Phoenix for that quick, one-half hour turn-around.

During-run procedure: The Air Traffic Controller and other experimenters were located in a control room adjacent to the simulator . Monitors available to the experimenter team were: 1. an X-Y plotter showing the aircraft path, 2. a visual scene display, 3. audio speakers, and 4. video screens showing views of the crewmembers and cockpit. The air traffic control and background conversations were scripted, although an occasional contingency intervention was required. Fuel available at the initiation of the approach to LAX was standardized by clearing the aircraft from an enroute hold when the fuel level reached 14,000 pounds.

Post-run procedure: After engine shut-down occurred it was reemphasized to the crew that they were not to discuss the flight until after the video peer review process. Each crewmember then rated, on 5-point scales, qualities of the simulator, the scenario, and study design as to: 1. the strength of agreement that the quality existed and 2. the essentiality of having the quality (for realistic crewmember performance). Twenty-one qualities were defined by statements like: "the simulator 'flies' like the actual aircraft" and "programmed anomalies (inserted problems) were realistic."

The crewmembers and observers then independently reviewed and commented on identical but separate video tapes of that crew's performance. Instructions to the reviewers solicited comments on events related to degraded or exemplary crew coordination and task performance. This video peer review process, and subsequent analysis, was adapted from an approach developed for analyzing clinical transactions in a medical setting (Frankel and Beckman, 1982). The day ended with the second phase of interviews related to how crewmembers represented tasks and perceived responsibilities.

RESULTS

Analysis of the data obtained during this study are in initial phases. The results reported herein are summary observations of the first author. Some major observations concern the decision-making process. A large number of crews did not initiate a timely or adequate assessment of their overall situation with respect to fuel and weather prior to their forced go-around at LAX. Many of these crews continued this trend while holding to perform a complete gear check procedure. Many stayed in the area for a second approach attempt at LAX despite considerable evidence of a deteriorating ceiling and RVR.

Since the weather situation did eventually prevent a legal and safe landing being made at LAX, a second go-around was required upon reaching the minimum decision height (conditions were reported as going from "minimums" to "below minimums" after an aircraft crossed the outer marker, thus permitting continuation of the descent to the minimum decision height).

A trend for many crews in exploring alternatives was to request weather information only for airports within the Los Angeles basin, ignoring, specifically, Palmdale, located nearby and experiencing good visual flight reference (VFR) weather. Although conditions at Ontario were approaching "minimums" early in the scenario, many crews did choose to land there. Often they did so with a remaining fuel level that would eventually commit them to Ontario despite further weather deterioration.

As could be predicted, considering the crewmember mix, considerable variability in performance, across crewmembers and crews appeared to exist. These differences did not, however, appear to be simple functions of recency on the line, or on the B707.

CONCLUSIONS

Although most conclusions await detailed data analyses, one major conclusion can be drawn from observations made during the study: performance, by many crews of a function called herein "critical situation assessment" was often inadequate. Usually, it seemed that a crew's overall performance could best be evaluated, or described, on the basis of how they performed three specific crew functions: crew coordination, flight task performance, and an ongoing critical assessment of the overall flight situation. The terms crew coordination and flight task performance are used here in their usual sense. The term "critical situation assessment" will refer to the ongoing process of critical assessment of flight and environmental factors that are, or may become, crucial to flight safety. In this scenario, the major, apparent factors of reference were fuel remaining, the weather condition and trend at the primary airport, the geographical location of the aircraft with respect to various possible alternates, and weather conditions and trends at these alternates.

Often it appeared that the above three major crew functions were fairly independent with respect to determinants of performance excellence. Thus, a crew might, for example, be well coordinated and perform control and navigation functions well, while failing to perform this critical situation assessment function. Figure 2 is a diagram of these three functions and their relationship with "resource management" as previously defined (Murphy, 1980): "The application of specialized skills to achieve a crew organization and process that effectually and efficiently utilizes available resources in attaining system objectives". Three sets of specialized skills were identified: 1. social and communication skills, 2. leadership and management skills, and 3. planning, problem solving, and decision-making skills.

A large amount of many kinds of data were obtained in this just-completed, full mission simulation. Results will be reported in a series of later papers.

Acknowledgements

The authors wish to express appreciation to the crewmembers and to Garth Knowlton and other Airline Training Institute personnel for their cooperation and diligence during the conduct of the study. Retired airline Captains Ted MacEachen, Jack Raabe, and John Scouten and Air Traffic Controller John Kurywchak were exemplary as members of the experimental team. Also thanks to NASA/Ames Research Center employees Drs. Clay Foushee and Al Lee for their many helpful suggestions and to Mark Allard, Marie Rider, and many others for their considerable assistance.

References

Blake, R.R. and Mouton, J.S.: *The new managerial grid*. Houston, TX:Gulf, 1978.

Calfee, R., Curley, R. and Siesfeld, A.: Communication scripts on the flight deck, In Hartzell, E.J. and Hart, S. (Eds); *papers from the 20th annual conference on manual control* NASA, Ames Research Center, 1984.

Foushee, H.C. and Manos, K.L.: Information transfer within the cockpit: Problems in intracockpit communications. Billings, C.E. and Cheaney, E.S. (Eds.), *Information transfer problems in the aviation system*. NASA TP-1875, 1981.

Frankel, R.M. and Beckman, H.B.; Impact: an interaction-based method for preserving and analyzing clinical transactions. In Pettigrew, L.S. (Ed), *Explorations in provider and patient interaction*. Humana, Inc., 1982.

Goguen, J., Linde, C., and Murphy, M.: Crew communication as a factor in aviation accidents. In Hartzell, E.J. and Hart, S. (Eds) *papers from the 20th annual conference on manual control* NASA, Ames Research Center, 1984.

Goguen, J. and Linde, C.: Linguistic methodology for the analysis of aviation accidents, NASA CR 3741, 1983.

Hackman, J.R. and Morris, C.G.: Group tasks, group interaction process, and group performance effectiveness: a review and proposed integration. In Berkowitz, L. (Ed.) *Advances in experimental social psychology*, Vol. 8, New York:Academic Press, 1975.

Murphy, M.R.: Analysis of eighty-four commercial aviation incidents: implications for a resource management approach to crew training. In: 1980 *proceedings annual reliability and maintainability symposium*, 1980.

Murphy, M.R.: Coordinated crew performance in commercial aircraft operations. In: *proceedings of the 21st Human Factors Society annual meeting*, 1977.

National Transportation Safety Board, Special Study: Flight crew coordination procedures in air carrier instrument landing system approach accidents, Report No. NTSB-AAS-76-5, 1976.

Randle, R.J., Jr., Tanner, T.A., Hamerman, J.A., and Showalter, T.H.: The use of total simulator training in transitioning air-carrier pilots: a field evaluation, NASA TM 81250, 1981.

Ruffell Smith, H.P.: A simulator study of the interaction of pilot workload with errors, vigilance, and decisions, NASA TM-78482, 1979.

Spence, J.T. and Helmreich, R.L.: *Masculinity and femininity: Their psychological dimensions, correlates, and antecedents*. Austin, Univ. of Texas Press, 1978.

ORIGINAL PAGE IS
OF POOR QUALITY

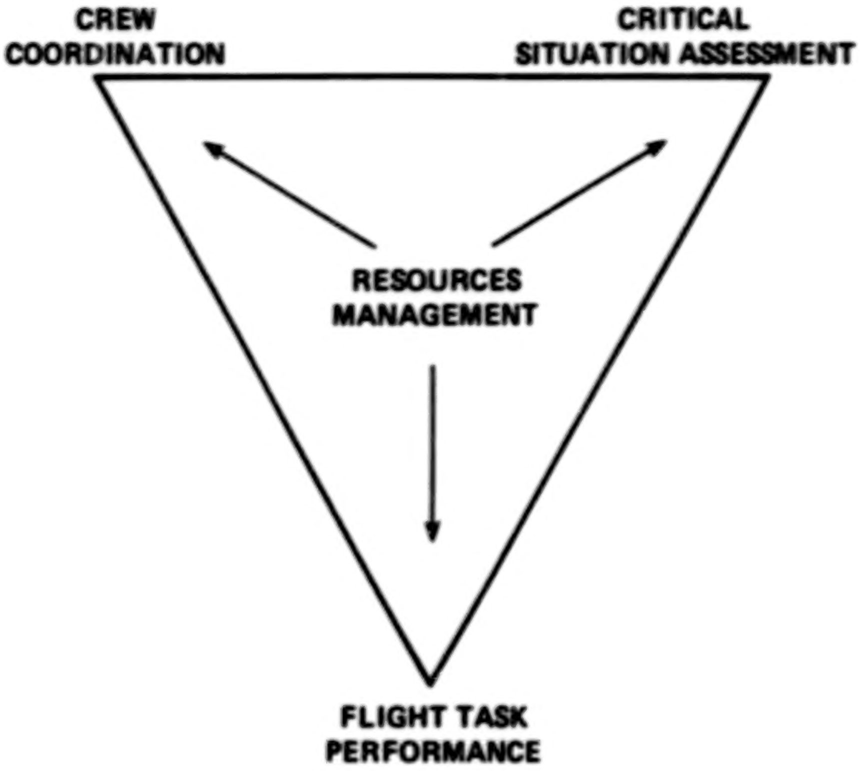

CREW COORDINATION
CRITICAL SITUATION ASSESSMENT
RESOURCES MANAGEMENT
FLIGHT TASK PERFORMANCE

Fig. 2: Major Crew Functions

TABLE I

RECORDED SIGNALS

ANALOG	DIGITAL
1. ILS/VOR DEV	1. MAIN GEAR ON GROUND
2. G/S DEV	2. LDG. GEAR WARNING LIGHT
3. TAS	3. AP ENGAGED
4. ROC	4. AP COUPLED
5. PITCH	5. AP IN G/S MAN
6. ROLL	6. NAV-1 ILS
7. HDG SIN	7. VIS SYST VALID
8. HDG COS	8. VIS SYST UNFADE
9. ALT (GROSS)	9. 1/4 VALVE OPEN
10. ALT SIN	10. LE FLAPS EXTEND
11. ALT COS	11. TE FLAP H>35
12. T/D LONG. ERROR	12. OBSERVER EVENT
13. T/D LAT. ERROR	13. OVER OUTER MARKER

N85
14557

UNCLAS

D22

COMMUNICATION ON THE FLIGHT DECK

To fly a modern jetliner it is necessary that well-trained flight crews work together in a coordinated way--they must be able to communicate about what they are doing. Their common understanding of what it takes to fly a plane serves this communication and coordination. However, evidence shows that accidents and incidents often are a result of failure to communicate (Billings & Cheaney, 1981, NASA Technical Paper 1875).

We approach the study of this problem from an information-prcessing point of view. We know that people's conceptions of what they are doing have an effect on how they perform a task and how they communicate about their performance of that task. We also know that if what they are doing is complex, they must find some relatively simple framework for representing it--otherwise the complexity overwhelms them.

(Specific examples - NTSB reports - AAR - 81-14,
AAR - 79 - 7; AAR - 80 - 14)

The first step, then, in studying this problem is to develop a model that reduces the complexity of flying a jet to a representation composed of a few relatively independent dimensions which capture the major features of this task. The model, once developed, should allow us to assess what crew members know (which is the basis for how they act), to look at actual performance, and perhaps to develop training recommendations.

The Model

Our preliminary model (Figure 1) is made of three dimensions. We think it is structured in the way a well-trained crew member might organize his understanding of the requirements of flying. The first dimension, flight tasks, represents the types of tasks crew members need to think about and perform during a flight. The second dimension is flight phases; each phase of a

PRECEDING PAGE BLANK NOT FILMED

flight has a somewhat different set of demands which crew members need to meet. Finally, the circumstances of the flight--whether events are normal or turning critical--demand different types of knowledge and attention from each crew member.

Here is a brief look at the composition of each of these dimensions:

Flight Tasks - Five categories make of this dimension:

-- External communication - the monitoring and management of all information exchanged with ATC, Weather Service, Company Dispatch, etc

-- Navigation - All activities involved in planning, directing, and monitoring the flight course

-- Vechicle Control - All activities involved in attaining the desired attitude and speed of the aircraft

-- Aircraft Systems Management - The operation, monitoring, and maintenance of all the systems, excluding the systems of vehicle control and communication

-- Transport Management - All activities undertaken specifically for the management of passengers and cargo

Each of these tasks needs to be handled during several phases of the flight, and each has a different bearing on the flight.

Flight Phases - This dimension is important because the division of a flight into phases is part of the thinking of every flight crew member. Their training and flight manuals tend to organize the technical responsibilities of crew members in this sequential fashion.

-- Preflight -ends when the plan begins to roll

-- Taxi/Takeoff - ends when the plane levels off at cruise altitude

-- Cruise - ends when the plane is cleared to descent

TASK CATEGORIES

PHASES	EXTERNAL COMMUNICATION	NAVIGATION	VEHICLE CONTROL	AIRCRAFT SYSTEMS MANAGEMENT	TRANSPORT MANAGEMENT
Preflight	discussions with ground crew, operations, dispatch, etc.				
Taxi/ takeoff		monitoring proper departure pattern, correct runway, etc.			
Cruise			maintain proper altitudes, modify air speed to balance weight, wind, schedule, etc.		
Descent/ approach				maintain proper pressure, monitor hydraulics, electric, and other systems	
Landing					give passengers departure info, connecting flights, gate, etc.

Figure 1. Activities under normal flight conditions

Siesfeld/Curley/Calfee
Stanford 6/84

-- Descent/Approach - ends with clearance to land

-- Landing

Arguably, there are many different ways to divide a flight into phases. We use these five because of the saliency of each phase's border event and because of differences in the relative importance of various activities within each of these phases.

Flight Circumstances - This dimension is best thought of as a continuum. At one end is the uneventful flight, and at the other end is the flight in crisis. We have cut the continuum into three segments; normal abnormal, and critical circumstances. These distinctions correspond closely with those made in most training programs.

The Interview

We used this model to build a structured interview that would let us assess how crew members think about flying a jetliner. Before I go on to describe how we collected data on flight crew knowledge, let me give a brief account of the technique of the structured interview.

It begins with a general question. For example, "Tell me how the flight engineer goes about doing his job." An individual with a well-formulated understanding of flying a jet might take a moment to "find" what he knows before giving an explanation. Any problems "finding" this information--in retrieving memories--should be easily resolved with a specific probe. "During a flight, what systems require the most attention from the flight engineer?" The interview continues as a series of general questions, backed with specific probes. It is structured in the sense that it is designed to match a well-defined conception of how knowledge should be organized.

In this study the interview was designed to match the structure of our model of the knowledge of flight.

This interview was given to each participant in the study. It consisted of two parts: the first was given the day crews were having differences training on the simulator; the follow-up was after the simulated mission and video review. We asked different types of question in each session.

During the first interview we asked that crew members answer the questions as they pertained to a normal, uneventful flight in which the captain was the pilot. We asked five types of questions (Table 1). One example, "Divide the sequence of events in a flight into four to six phases. As you do so, list the event that marks the border between each phase."

For the follow-up, we concentrated on the interplay between flight tasks and flight circumstances. Again we asked these types of questions (Table 1). An example here, "In your experience, what are the three most common occurrences which disrupt activities related to vehicle control? What would the immediate consequence of each occurrence be?"

Findings

We have completed the interviews, and have begun the data analysis. We will compare the findings with performance measures taken during the simulation. As we do this, we expect to find more effective coordination among crews with members who

1. share a common understanding of activities and responsibilities
2. are most aware of the ongoing activities of the other crew members
3. are able to analyze the demands of the flight from the dimensions of task categories, and flight phases
4. are sensitive to factors affecting task performance during abnormal and emergency conditions.

Table 1

Interview Topics

PRE-SIMULATION

1. Job descriptions: major activities of each crew member
2. Flight sequence: division of flight into major segments
3. Flight tasks: division of flight into task categories
4. Specific activities: position x phase x task descriptions
5. Training background: abnormal + emergency procedures; crew coordination + communication

POST-SIMULATION

1. Abnormal events: frequency + effects of most common abnormalities in each task category
2. Emergency events: frequency + effects of most serious emergencies in each task category
3. Scenario response: expectations for crew response to an abnormal and an emergency scenario
4. Simulation recall: recall of crew activities during three points of the simulation

---

Siesfeld/Curley/Calfee
Stanford 6/84

Although we have just begun this analysis, we developed some impressions about the knowledge structures of crew members while collecting these data. These first impressions are leading us to begin to reformulate the model. After presenting these impressions, I'll finish by describing our initial reformulation.

Impressions

1. From the first interview, it is evident that the dimension of flight phases makes a lot of sense to crew members. In fact, members kept giving us finer divisions of flight phases than we identified. The division they made most consistently, that differed from ours, was in the preflight. Where we had one, they saw two phases. They identified preflight-planning, all the activities undertaken before entering the plane, and preflight-operations, those activities associated with preflighting the plane.

2. Also, from the first interview we got a different picture about the dimensions of task categories. It seems that most crew members organize this dimension by describing what different crew positions are responsible for during different phases. Only a few crew members generated a list of task categories that covered all the tasks that need to be performed during the course of the flight.

3. The last impression comes from the second interview. It appears that each crew member has a different notion of what constitutes an abnormal and a critical situation. Because of this, certain situations are viewed and handled in various ways.

Revised Model

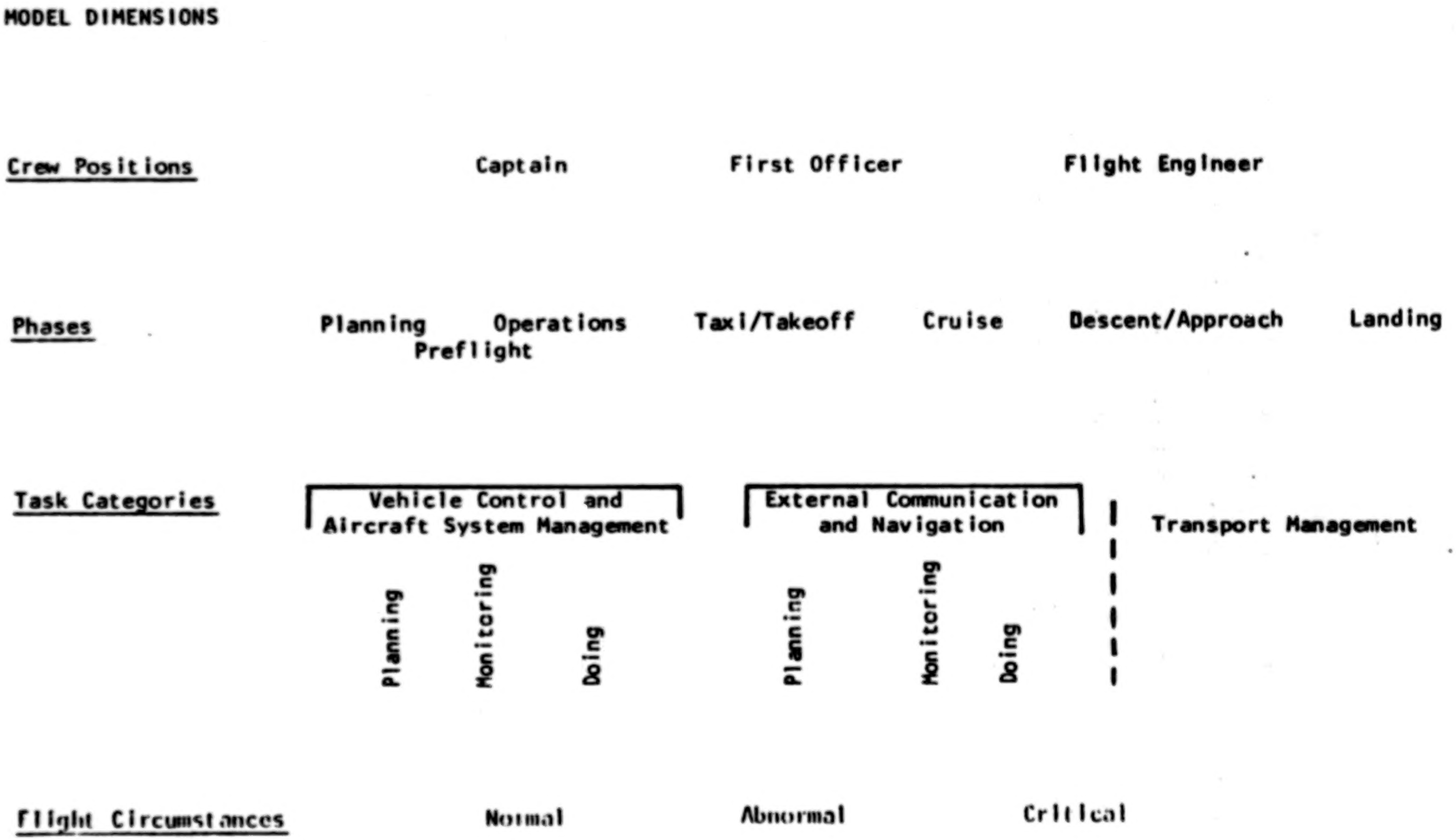
These impressions are leading us to reconsider our model. The major dimensions are sound, but some of them need to be reorganized. Figure 2 shows the changes we have made.

1. We have added the dimension of crew position. This was implicit in our first model. However, because crew members often decide what they will do based on their assumptions of what others will do, it is important to make this dimension explicit.

2. We have split the phase of preflight into -planning and -operations. This reflects the distinction that most crew members made.

3. Finally the five task categories can be reorganized into three. Vehicle control and aircraft system management came together into a category best described as keeping the plane aloft. External communication and navigation combine to get the plane from point A to point B. Transport management stays as it is, but is de-emphasized. Managing passengers and cargo is not essential to flying the plane (however, there would be no flying without passengers or cargo).

Each of the new categories is composed of three types of tasks. These are planning, monitoring, and performing tasks associated with each.



MODEL DIMENSIONS

Crew Positions: Captain, First Officer, Flight Engineer

Phases: Planning Preflight, Operations, Taxi/Takeoff, Cruise, Descent/Approach, Landing

Task Categories:

- Vehicle Control and Aircraft System Management
  - Planning
  - Monitoring
  - Doing
- External Communication and Navigation
  - Planning
  - Monitoring
  - Doing
- Transport Management

Flight Circumstances: Normal, Abnormal, Critical

Figure 2. Results: Revised model

Siesfeld/Curley/Calfee
Stanford 6/84

Conclusion

I will conclude with the following comments about this research. We are aiming to set the groundwork for a detailed picture of how crew members think about their jobs. This has immediate implications for research in this field--in the development of instruments to assess crew communication, coordination, and decision-making. And if productive, it has long-term implications for how flight crew training should be conducted.

Training

N85
14558

UNCLAS

N85 14558 D23

DETERMINING TRAINING DEVICE REQUIREMENTS IN ARMY AVIATION SYSTEMS

Captain Michael L. Poumade
Directorate of Training and Doctrine
US Army Aviation Center
Fort Rucker, AL 36362

The views, opinions, and/or findings contained in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other official documentation.

ABSTRACT

The growing complexity of Army Aviation systems has been accompanied by increasing sophistication and cost of training devices for those systems. Conversely, some less costly, part-task trainers have also been proposed to address some of the aviation training requirement. These trends have caused the decision makers to make trade-offs in order to provide the most training effective system, while staying within constrained resources. A method of quantifying and facilitating the decision making process is clearly needed.

This paper discusses such a methodology which applies the systems approach to the training problem. Training is viewed as a total system instead of a collection of individual devices and unrelated techniques. The core of the methodology is the use of optimization techniques such as the transportation algorithm and multi-objective goal programming with training task and training device specific data. The role of computers, especially automated data bases and computer simulation models, in the development of training programs is also discussed.

The approach described in this paper can provide significant training enhancement and cost savings over the more traditional, intuitive form of training development and device requirements process. While given from an aviation perspective, the methodology is equally applicable to other training development efforts.

PURPOSE

The purpose of this paper is to describe a methodology which applies the systems approach to aviation training device development. Training is viewed as a total system and not merely a collection of individual devices and unrelated techniques. Device development is portrayed as both a training enhancement and a cost avoidance measure. Automation and standard operations research techniques are integrated into the methodology as a foundation of the decision making process.

Although presented from an aviation training perspective, this methodology is equally applicable to training device development efforts in other areas.

It is felt that this methodology can aid decision makers, offer more comprehensive training programs and provide monetary savings over the usual training device development process. It is also hoped that this paper will stimulate thought and discussion among users, trainers, analysts and decision makers which will result in improved methodologies and more training and cost effective training systems.

BACKGROUND

Flight simulators have a long history. The first flight simulators are said to have been introduced about 1910, with the Army Air Corps procuring six devices in 1934 to train pilots for mail routes. The training utility of these systems was recognized and over 500,000 airmen received training in 10,000 link trainers during World War Two.

Technological advances were made in synthetic flight simulators with the introduction of the UH-1 Flight Simulator. That simulator, currently the most numerous in the Army, was primarily intended for instrument flight training. The CH-47, AH-1 and UH-60 Flight Simulators incorporated visual systems which greatly expanded the training capability of the simulator. The AH-64 Combat Mission Simulator (CMS) is a further expansion of training device capability with interactive threat and other features.

Parallelling the increased use of flight simulators, an increase in the employment of part-task trainers has also occurred. These range from cockpit procedure trainers (mock-ups of the cockpit with operable instruments and capable of fault insertion to train pilots in routine procedures such as engine run-up, shutdown and emergency procedures); to Cockpit Weapons Emergency Procedures Trainer, or CWEPT, which approaches the capability of a flight simulator but lacks motion; to computer aided instruction in the form of classroom systems trainers or desktop type trainers oriented toward specific groups of training tasks.

All of these devices cost money, as a total training system hundreds of millions of dollars are involved. In 1982, the Army Audit Agency (AAA) valued flight simulators alone at about 616 million dollars. Naturally, programs of this size receive attention from several sources. Competitive funding is a fact of life and decision makers at all levels must conduct trade-offs among various programs to reduce expenditures while providing an acceptable product. In the aviation community the desired product is a fully qualified aviator. The most cost and training effective means of producing that aviator is the goal of the decision makers concerned. The increasing availability of computer resources and trained analysts have enabled the aviation training community to quantify the training requirement as never before possible. This new methodology is necessary to aid the decision maker in making his trade-offs by answering some essential questions, such as:

- What are the characteristics required for a given device? Why?
- What tasks will be trained on this device?
- Why is the quantity of this device required?
- Why is there a variety of device types required for this system?
- What are the best (training and cost effective) locations for these devices?

Some military facts of life complicate the answers to these questions.

- System characteristics (hardware and operational) are evolutionary; also they may not be standard. (e.g., the AH-1 COBRA has had 6 models since 1967, 5 currently in service with the total Army and other models with the Marines)
- Training technology is also evolutionary.
- Unit/aircraft locations are dynamic, not static.
- The aviator population is diverse.
- The training environment encompasses other systems (e.g., combined arms, threat).

The methodology about to be presented does not imply that aviation training has not been studied previously. On the contrary, several studies by the Army, by other services and by contractors have been conducted over the years. Unfortunately, many of these studies have been either questioned, rejected or soon became outdated. A number of factors may have contributed to those problems, resource and methodology constraints were often involved. The major deficiencies could be summarized as follows:

- Task analysis not comprehensive.
- Device analysis restricted to synthetic flight simulator.
- Task training requirement not identified by time and iterations.
- Data base not established or not used.
- Sensitivity analyses not performed.
- Training and device strategies determined consensually without access to detailed, quantitative information.

METHODOLOGY

The program which is presented now considers the questions raised by decision makers and the shortcomings of earlier studies. The approach incorporates several features from those studies and from proposed studies. It also adds some rather standard operations research techniques to help quantify the device requirements.

The process for determining and quantifying training devices for Army aviation systems can be viewed as consisting of eight major steps. These steps are shown in figure 1.

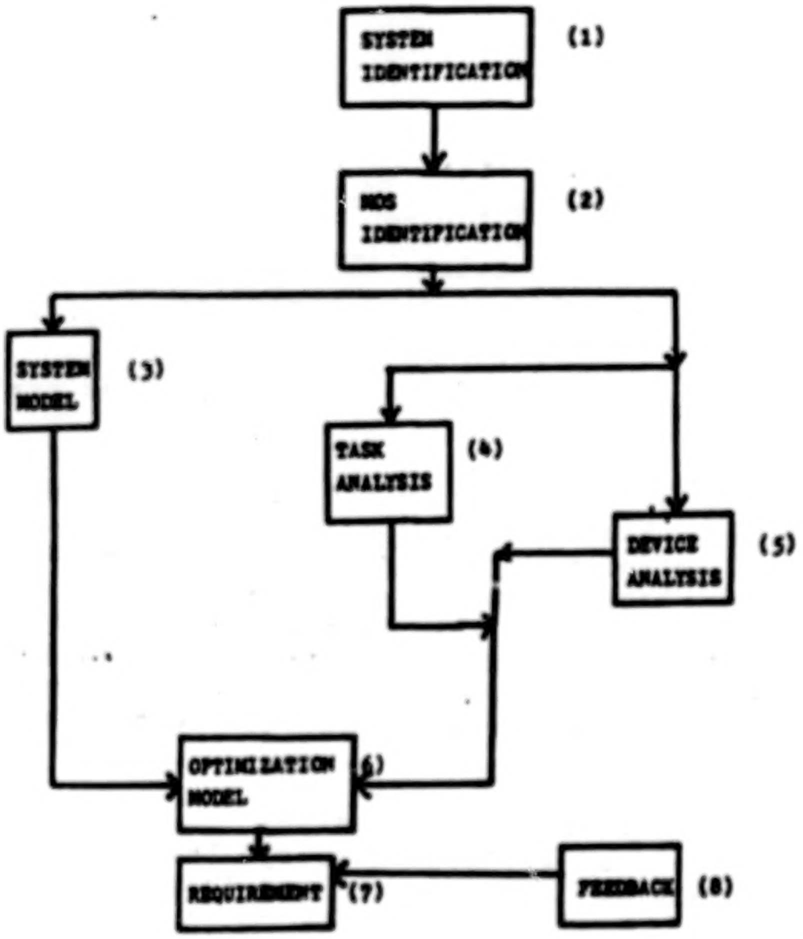

SYSTEM IDENTIFICATION
(1)
MOS IDENTIFICATION
(2)
SYSTEM MODEL
(3)
TASK ANALYSIS
(4)
DEVICE ANALYSIS
(5)
OPTIMIZATION MODEL
(6)
REQUIREMENT
(7)
FEEDBACK
(8)

FIGURE 1. GENERALIZED METHODOLOGY FLOW

STEPS 1 AND 2 SYSTEM AND MOS IDENTIFICATION

Steps 1 and 2 are performed by the decision maker. Although they may seem trivial, they are crucial for problem formulation, scope, data needs and feasiblity analysis. For instance, is the decision maker interested in helicopter training or in attack helicopter training? If attack helicopters, are both the AH-64 and the AH-1 to considered? If only the AH-1, are all models to be analysed or only one. What military occupational specialties are to be considered? Warrant Officer pilots? Commissioned pilots? Maintenance personnel? Armament personnel? etc.

Steps 3, 4, and 5 can be conducted concurrently if resources allow.

STEP 3 SYSTEM SIMULATION

Step 3 involves identification and modelling of the training system, especially the manpower pool. Relationships must be explored to allow training needs in future years to be forecast. A simulation of the training system over time can then be performed to identify student populations requiring device use at the institution and aviator populations requiring devices for sustainment. It may be possible to forecast the requirements of competing or companion systems as well, thus identifying potential problems with ranges or training areas. The training system's sensitivity to changes in aircraft production, doctrinal staffing ratios of crews to aircraft, training course lengths, etc., can also be examined.

STEP 4 TASK ANALYSIS

Step 4 must identify all tasks, not merely device specific tasks. Task data gathered can then be included in a data base for use in the future should the training system change. The task analysis should include a comparative analysis of tasks for other systems, this is one of the major ways to begin training development for a system just being conceptualized. When a comparative cnalysis is used the task analysis must be repeated as the system matures and empirical data becomes available through operational tests. Data items include task iterations required for proficiency, time per iteration and functional training characteristics (FTC) associated with the task. Generation of a basic task list can be aided by automation. An elementary means is to consolidate existing task lists for aircraft of interest and code them by type into a file. This enables a retrival of attack tasks, cargo tasks, etc. A more sophisticated technique is illustrated by the Computer Aided System for Developing Aircrew Training (CASDAT), an interactive computer program developed for the US Navy. This program creates a task list by asking the developer a series of questions relating to aircraft performance, mission and crew configuration and then retrieving the appropriate tasks from its data base.

STEP 5 DEVICE ANALYSIS

Step 5 includes an analysis of existing and potential devices. Data items include cost, reliability and availability data, and functional training characteristics which the device is capable of replicating.

It is imperative that the information gathered in steps 3-5 be entered in to a data base(s). This allows a ready means of retrival and comparison. It also allows for a rapid and comprehensive update procedure as the aircraft or training system evolve.

STEP 6 OPTIMIZATION MODELS

Filter and Sort Routines

Step 6 involves further application of operations research and automation techniques. The product of the task and device analyses can be merged through a program which can identify the capability of each device to train each task based on assumed or demonstrated functional training characteristics. The program is basically a sorting routine. If voids exist where no device is capable of training a task based on FTCs, then those unsatisfied functional training characteristics may constitute a new device design requirement. The requirement may be added to an existing device or a separate device may be conceived. The task list itself may be filtered after determining device capability to reduce the list based on such considerations as previously learned tasks or tasks that are trained incidental to normal missions, etc. It is important that device-task capability be established first because changes in entry level characteristics or operational policies may change what tasks are previously learned or routinely performed. Also redundant training capabilities among candidate devices can be identified to perhaps reduce device variety.

The use of a computer program allows several excursions to investigate the impact of changing device or task FTCs, task lists or device availability.

Transportation Algorithm

When device-task capabilities are known (or assumed) they can be used in an optimization program. If linear relationships can be assumed for device operating costs and task training times, then a linear program such as the transportation algorithm can be used. This program is a streamlined, more computationally efficient version of the SIMPLEX method. Several texts can be found on this technique. The Transportation Algorithm is so named because of the type of problem it was originally used to solve. The problem was formulated to find the minimum cost of transporting a given supply of goods from a certain number of sources to satisfy a given demand at certain destinations, the transportation cost between a source and a destination was given as well as any other key relationships. For the training application the basic requirements are: a cost matrix, expressed in standard units such as dollars per minute; a device-task capability matrix, this can be of the form of 1-0 for capable or not capable, if training effectiveness ratios are available they can be used: a set of constraints indicating the training time required per task; and another set of constraints indicating the time available per device. The cost matrix represents the operating cost of the device and corresponds to the transportation cost. The device-task capabilities correspond to key relationships between sources and destinations, in this case the sources are devices and the destinations are training tasks. The supply is the training time available and the demand is the training required. The program assigns training iterations to each device subject to the devices capability, time available and task requirements remaining. After each assignment the task time and device time are decremented and the training cost is calculated. After all assignments possible are made, the computer repeats the process attempting to modify the assignments so that a less expensive training program results while still satisfying all the constraints.

Of course, this model has some rather extensive data requirements and it is a bottom-up approach to training requirements. The reliance on steps 4 and 5 are clear. Task data (times, iterations, iteration durations, etc.) can be obtained in numerous ways. For emerging systems a comparative analysis or analogy to existing systems can provide a first cut. The methodology should be repeated (see step 8) as data is refined. Refinement can be done with subject matter expert inputs (a modified consensual delphi technique is one possibility); developmental/operational tests; follow on evaluations; and logistic support analysis reports (LSAR) are other means. While the accuracy of the model is dependent upon the input, input in this case is defined, can be replicated and can be easily updated - all failings of earlier, more intuitive approaches. The data requirements are not show stoppers. If models of this type were consistently applied to training systems much data could be already available for use during the concept formulation of new systems, due to task commonality.

The output of the above model can be structured to identify unsatisfied training requirements, training time and cost per task, training time and cost per device, unused training time per device, and total training costs. This allows the decision maker to quickly grasp the impact of training certain tasks or the impact of a certain device's availability. Again sensitivity of the training program can be easily investigated - the analyst can run the model with different training requirements.

The above model will produce the optimal training strategy per aviator, the results of the training system simulation in step 3 when combined with the above product can identify the total device requirement over time for the system and the assumptions used. The time factor is important since it presents potential milestones for device acquisitions. The total quantities are also important, as economies of scale may influence the contract and budgeting processes.

Multi-Objective Goal Programming

One more model that can be used to fine tune the requirement process is the multi-objective goal programming model described by Dr. Ignizzio in several texts. This model would allow refinement of the device mix based upon more considerations than training effectiveness and operating cost. A series of priorities which might include operating costs, acquisition costs, number of tasks trained (or not trained), availability and reliability data, and variety of devices could be used to modify the devices selected. The output of this model might require that the transportation model be repeated, especially if the variety of devices were altered.

Basis of Issue Plans

When the required device mix is established, the next need is to determine the optimal basis of issue plan (BOIP) or stationing plan for the devices. Some devices would have been designed for use at the institution or for use at each unit. Their requirement would have been either justified or modified by the procedures above. For those devices (such as flight simulators) that must still be assigned the transportation algorithm can be modified to determine the optimal plan. In this case one needs to know the aviator densities at each location and travel costs between each location. From the aviator densities and individual training requirements, the training requirement per installation can be determined. The appropriate information for the model is the training per installation, travel cost between installations, capability of an individual to train at an installation (e.g. long range force structure plans may rule out positioning a simulator at a certain site) and simulator time available. This program would produce an optimal plan for the assumptions made. It would enable sensitivities such as force structure changes to be easily evaluated.

STEP 7 REQUIREMENT

All of the above steps now feed into step 7, the actual statement of the requirement and its funding. The steps above should provide adequate justification for the requirement and insights into training and cost implications if the requirement is not satisfied. Additionally, the use of automated data bases and in-house models will allow the training developer to respond to the questions asked by various decision makers throughout the budget process. The application of judgmental factors, subjective and objective, may alter the requirement. If the requirement is altered, the mechanism is in place to assess the impact and modify the training plan and system accordingly.

STEP 8 FEEDBACK

The last step in the requirement process is feedback. As a system matures more data will become available. Tests may reveal that a device does not meet expectations, or that it exceeds them in regards to performance, cost or delivery schedule. Tests can also provide empirical data if only subjective estimates were previously available. Doctrine may change causing the addition of new training tasks or the elimination of old ones, the same is true for hardware developments. Force structure may change, or the characteristics of the training population may become different. Any of these changes may occur, and it is nearly certain that at least some will. Each change will in turn affect the training system. The mechanism to evaluate the change will be in place, only requiring that new data replace the old and the techniques be repeated.

Feedback Systems

Many feedback systems already exist for Army aviation systems. Most notably, they include: the US Army Safety Center, which tracks accident data; the Directorate of Evaluation and Standardization, which records the results of assistance and inspection visits to aviation units worldwide; the Directorate of Training and Doctrine, which tracks the results of the Annual Aviator Written Examination; the Device and System Training Information System, a data base on Army training devices maintained by the US Army Training Support Center (ATSC); the Training Development Information System, also at ATSC, which is a data base containing information on tasks for all MOSs; and finally the US Army Soldier Support Center, which maintains demographic data on the Army training population.

While the procedures identified in step 6 were linear methods, one should note that the refinement of data may lead to the use of nonlinear methods to determine the optimal training device requirments. This is especially true when the cost, learning and learning decay curves can be defined. While the application of linear optimization methods have been demonstrated at the US Army Aviation Center in the training context, the nonlinear techniques remain to be pursued.

SUMMARY

The methodology presented requires early involvement of the decision maker and early definition of the training system. It relies on simulation to forecast the training requirement as driven by the training population and by competition with other systems. It emphasizes analyses of all tasks and all devices and the creation of automated files for use in updating training requirements. The approach also utilizes computer driven optimization models, building on the results of earlier steps, to determine optimal training device mixes, quantities and locations. These steps assure that a systems approach is followed, the relationships among all tasks and devices, as well as the external training environment are considered. The implementation of computer methods enable the analyst or decision maker to easily evaluate system sensitivities and respond to trends indicated by existant feedback systems.

CONCLUSION

Although no single element of the methodology, by itself, is original, the combination of steps and subelements is not commonly found in the training development process. The need for quantative methods is urgent; the techniques are there. It only requires that they be implemented. The methodology above is a road map for that implementation. Resource constraints such as time, personnel or money may preclude its full implementation. Even so, short cuts and side excursions from this road map may lead to the final goal of supportable and effective training device requirements.

REFERENCES

The author wishes to acknowledge the influence and contributions made to this methodology by the following individuals: Dr. Erby Fischer; Dr. George Huntley; Dr. Jack McCracken; and Ms. Maryann Shipley, all currently or formerly with the Directorate of Training and Doctrine, Fort Rucker, AL.

1. Bobby R. Adams, "The Synthetic Flight Training System Program", p. 24, ARMY AVIATION, Volume 32, Number 9, September 30, 1983.

2. Jesse Orlansky, "The Name Of the Game Is Saving $", p. 75, ARMY AVIATION, Volume 32, Number 9, September 30, 1983.

3. "Audit of Synthetic Flight Training System", US Army Audit Agency, Audit Report: SO 82-6, March 1, 1982.

4. James P. Ignizzio, "Linear Programming In Single & Multiple Objective Systems", Prentice Hall, Inc., Englewood Cliffs, N.J., 1982.

N85
14559

UNCLAS

N85 14559 D24

THE DESIGN AND USE OF SUBTASKS IN PART TRAINING AND THEIR RELATIONSHIP TO THE WHOLE TASK

Amir M. Mané, Michael G.H. Coles, Demetrios Karis,
David Strayer and Emanuel Donchin
Cognitive Psychophysiology Laboratory
University of Illinois

INTRODUCTION

The issue of part versus whole training has attracted researchers from the early days of experimental psychology. Common sense dictates that a massive body of knowledge should not be taught as a whole (Adams 1960). For the same reason, if a part of the task is very difficult, one should not go over the task in its entirety. Repetitive training on the difficult part should lead to better results (Seymour, 1954). Several part training methods have been developed, including pure part, progressive part, repetitive part, retrogressive and isolated parts (Stammers and Patrick 1975). Typically the scientific question was whether or not training on "parts" of a task is a beneficial enterprise. This question cannot be answered without a prior determination of the way the task will be disassembled for the "part" training. The major controversies in this area have concerned the manner in which tasks are decomposed rather than the specific effectiveness of part-training, (Adams, 1960; Annett & Kay, 1956; Briggs & Naylor, 1962). Furthermore, Naylor and Briggs (1963) have argued persuasively that the effectiveness of part-training depends on the degree to which a task is decomposable.

While the importance of decomposition appears self-evident, investigators are confronted with a major hurdle. There are currently no consensual, objective, techniques for effecting such a decomposition. Much of what passes for task-analysis is essentially intuitive. The most commonly used techniques (e.g. time-line analysis) are very descriptive. It is difficult to infer any relation between the resultant components and the elements of a model of the cognitive structure of the operator. That is, there is little that relates the task components to aspects of human skills and cognitive resources. Other attempts (Miller, 1967; Gagne, 1970) have been made to discuss the skill in psychological terms. However, these attempts are purely descriptive and are not tested against any objective criterion.

In a previous report (Mané, Coles, Wickens and Donchin, 1983) we described an attempt to apply a decomposition methodology developed by Sternberg (1969) in the domain of mental chronometry to the analysis of complex tasks. Sternberg's approach assumes that if the effects of two independent variables are additive the two must affect independent aspects of the information processing system. Two variables whose effects on performance interact are viewed as affecting the same aspect. We applied this methodology to study the structure of a computer-controlled video game which was developed for research purposes and was named "Space Fortress". In the first study we manipulated the difficulty of the game along several different dimensions. The results indicated that the Space Fortress game is a complex task, whose successful performance depends on at least three skills: appraisal, motor, and perceptual-motor. These skills, and the associated aspects of the task that draw upon them, can be isolated. In terms of training, the data implied that a successful training regime should include two parts. First, part training on appraisal, motor, and perceptual-motor skills. Second, whole training with gradual increase in difficulty of that component of the task which draws on perceptual-motor ability. Note that, because we found an interaction between perceptual-motor and other abilities, a regime that only included part training could not be expected to succeed. Substantial whole training in which all skills are integrated is clearly necessary.

Based on the task structure we developed sub-tasks that (a) depend for their successful performance on the isolated skills and (b) provide training whereby these skills can be developed. If our subtasks meet there criteria, performance on a particular sub-task should be related to specific aspects of the whole task performance, the same aspects which were tied to the isolated skill. If these relationships are obtained, they would provide converging evidence for the alleged task structure.

Game description. In the Space Fortress Game the subject was seated in front of a display unit on which a number of elements were presented (see figure 1). His task was to destroy a Space Fortress (Fort), located in the center of the display, by pointing his space ship (ship) at the fort and firing missiles at it. To destroy the fort, the subject had to first hit the fort with ten single shots, before firing a burst of two shots with a maximum inter-shot interval of 250 msec. The number of single hits on the fort was displayed at all times by a digit located beside the fort. The subject controlled his ship and fired missiles using a standard aviation control stick manipulated by the right hand. When the trigger of the stick was depressed, missiles were fired from the ship in the direction in which the ship was pointing. Forward movements of the stick caused the ship to accelerate. Lateral movements caused the ship to rotate. Because the ship was flying in a frictionless environment, it continued to fly in the direction in which it was pointing unless it was rotated and thrust was applied. Thus, control of the ship was a complex perceptual-motor task. In trying to destroy the fort, the subject had to deal with a number of different obstacles. First, the fort could rotate, "lock-on", and fire shells at the subject's ship. Thus, the subject could not remain stationary. Second, from time to time mines emerged from the fort and chased the subject's ship. Every missile which was fired when a mine was present on the screen was ineffective against the fort. Mines could be of two types, "friend" or "foe". A letter, presented in the center of the screen, was used to designate the type of the mine. As part of the instructions prior to the experiment, the subject was told which letters identified the "foe" mines. The subject had to act differently depending on the mine type. If the mine was a "foe", the subject had to first identify it as such before firing a missile to destroy it. Identification was accomplished by the depression of a button located on top of the joystick. The subject had to depress this button twice, with a prescribed interval between the two presses to accomplish identification. If the mine was a "friend", no identification response was required. Hitting the mine with a single missile shot "energized" the mine, i.e. the mine changed direction and moved rapidly in the direction of the fort. The identification procedure was presented to the subjects in terms of selection of a weapon system according to target. If the subject failed to destroy a "foe" mine or to energize a "friend" mine within 10 sec, the mine self destructed. The interval between mine appearances was 4 sec, and the subject had to fire at the fort during that interval.

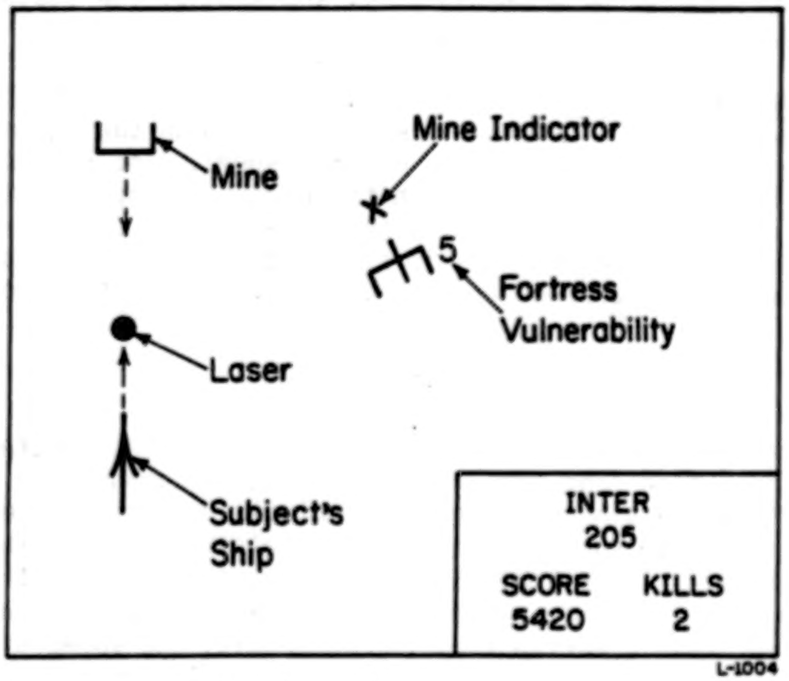


Figure 1 The elements of the Space Fortress game

THE DEVELOPMENT OF THE SUBTASKS.

Four subtasks were designed on the basis of the task analysis. The four subtasks were: (a) production of the double button press; (b) recognition of the letters designated for a friend or a foe; (c) aiming of the ship; and (d) control of the ship movement. These tasks depend on skills related respectively to (a) motor, (b) appraisal, and (c,d) perceptual-motor processes.

Aiming sub-task. In this task the space ship can only rotate. The subject had to aim and shoot at a mine which was stationary. The subject's score reflected the number of mines destroyed in a 2 min block.

Double-press sub-task. In this sub-task the letter 'X' appeared on the screen. The subject was instructed to respond by a double button press. The second press had to be executed 225 msec following the first, but any response interval between 150 and 300 milliseconds was considered correct. A subject's score reflected the number of times that he correctly produced the double press.

Mine-identification sub-task. The subject was taught the letters which identified a foe. A letter was presented (designating either friend or foe). The subject was asked to identify the letter and respond accordingly by pulling the trigger or by a double press and a shot. The subject received points for each correct identification.

Ship-control sub-task. In this task the subject had to learn how to slow down the ship. The ship was presented alone on the screen. Then, the program accelerated the ship to its maximal speed. The subject first had to rotate the ship so that it pointed in the direction opposite the one in which it was heading. Then, he had to use the acceleration control to bring the ship to a slow speed. When the ship slowed down to a criterion speed, the subject's score was increased and the computer accelerated the ship again.

PROCEDURE

Forty male, right-handed subjects with normal or corrected-to-normal vision were recruited from the university student community. The schedule used for subtask training included three blocks of aiming (2 min each), one block of button press subtask (2 min.); one block of letter subtask (2 min) and two blocks of ship control (5 min each). After subtask training, subjects received training on the whole task according to an adaptive algorithm. For this adaptive training procedure, the speed of the hostile elements (i.e. mines and fort shells) was initially set to a slow speed (5). Then, as the subject's performance improved (defined by survival time and fortress destruction rate), the speed of these elements was increased.

Training in this adaptive mode proceeded for a total of 20 five minute blocks given over three sessions. For the last block of each session, the 5th, 12th, and 20th blocks, the speed of the hostile elements was fixed at the maximum speed of 20. These blocks allow us to compare the progress of the subjects on identical conditions.

RESULTS

The first step in our analysis concentrated on the reliability of the subtasks scores. The correlations among the three administrations of the aiming task were 0.78; 0.79; 0.83. The correlation between the two blocks of ship control was 0.92. The correlations between the three aiming blocks and the ship control blocks ranged betweeen 0.54 to 0.64. The double press and the letter recognition subtasks did not correlate with each other nor with the other subtasks.

Some of the correlations between the scores of subjects in the subtasks and their scores in the whole task are presented in figure 2. All three blocks of aiming and the two blocks of ship control correlated with success of the subject in the task at the three probe points. The correlation of third block of aiming with score on the fifth block was 0.54 (unless specified otherwise p is always < 0.01). The correlation was higher with score on the 12th and 20th blocks: 0.70 and 0.72 respectively. The same is true about the ship control subtask which correlated at 0.38 ($p<.05$) 0.72 and 0.69 with the three blocks respectively. It appears that the correlation of performance on both subtasks with whole task performance increases as training proceeds. In comparison, the correlations between the subjects' scores on the first block of whole training (in the adaptive mode) and his scores on the 5th, 12th, and 20th blocks were 0.64, 0.62 and 0.60 respectively. The other two subtasks did not correlate with the subjects score on the whole task.

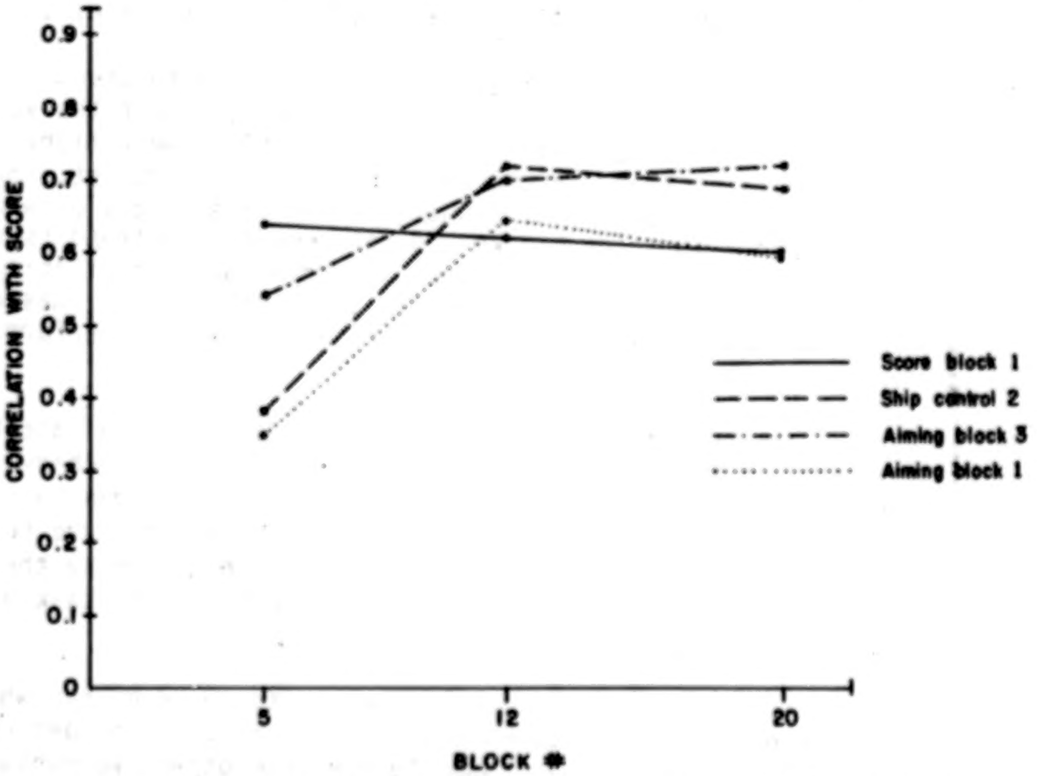


Figure 2 The correlations between scores of the subtasks and success in the whole task on blocks 5 12 and 20.

The same pattern of results appears for a wide range of measures of performance in the whole task. As an example, the number of times that the subject hit the fort correlated with the aiming subtask at 0.59, 0.68, 0,70 at the end of the three sessions. For the same blocks ship control correlated at 0.57, 0.69 and 0.78. Another example is the movement through hyperspace. Subjects were instructed to avoid moving from one end of the screen to the other because of the negative consequences which followed such a move. The inverse of the number of times that the subject moved through hyperspace correlated with aiming at 0.48, 0.59, 0.71 and with ship control at 0.60, 0.58 0.80 at the end of the three sessions. Once again, the other subtasks did not correlate consistently with any of the whole task measures.

CONCLUSIONS

The aiming sub-task, which represents a basic element of piloting, and the ship-control sub-task, which taps a more complex element of piloting - the ability to slow down - are both linked to the success of the subject in the task as a whole. A practical use of that finding is that performance on these tasks may be used to predict the subject's eventual performance after completion of training. The prediction is powerful if we consider the fact that it is based on a very simple version of the task, and can be obtained with a minimum investment of time. Indeed, the aiming subtask, which takes a total of six minutes was successfully used as a screening and placement test in another experiment which utilized the Space Fortress task (Mane, 1984).

It is interesting to note the changes in the correlations among variables along time. The correlations of the two subtasks with the various indicators of success increased. That is, the correlations were higher for the late blocks of training than for the early blocks. In contrast, the correlation between performance of the subject in the first block of whole training with performance in later blocks did not show such a trend (see figure 2). A possible interpretation is that the importance of piloting of the ship grows with the development of the skill. In other words, mastery of that skill determines the ability of the subject to improve his performance with other elements of the task.

The results of the reported experiment cannot be used to evaluate the overall effectiveness of the part training method. This can be achieved only with a transfer of training experiment (See Mané, 1984 for the evaluation of the effectiveness of part training). However, the findings can serve to contrast the different subtasks, and to evaluate their relevance to the performance of the task as a whole or to specific aspects of the task at different stages in training.

Overall, there is a high correlation between performance on the whole task and performance on the two subtasks that were related to the perceptual-motor skill. No such correlation was found between the other two subtasks and the performance of the whole task. These correlations indicate that the skill tapped by the perceptual-motor subtasks is central to the performance of the task.

There is an apparent conflict between the fact that in the task decomposition there were three skills which constituted the whole task and in the present analysis one skill dominates. However, the game in its current version was performed at the easy levels of the appraisal procedure and motor response. This deviation from the prior version may be the reason for this conflict.

References

Adams, J. A. Part trainers. In G. Finch (Ed.) *Educational and Training Media: A symposium.* Washington, D. C.: National Academy of Sciences, National Research Council 1960.

Annett, J., & Kay, H. Skilled performance. *Occupational Psychology*, 1956, *30*, 112-117.

Briggs G. E., & Naylor, S. C. The relative efficiency of several training methods as a function of transfer task complexity. *Journal of Experimental Psychology*, 1962, *64*, 505-512.

Gagne, R. M. *The conditons of learning*(2nd ed). New York: Holt, Reinhart and winston, 1970.

Mané, A. M. Adaptive and part-whole training in the acquisition of a complex perceptual-motor skill. Submitted to the Proceedings of the Human Factors Society - 28th Annual Meeting-1984.

Mané, A. M., Coles, M. G. H., Wickens, C. D., & Donchin, E. D. The use of the additive factors methodology in the analysis of a complex task. Proceedings of the Human Factors Society - 27th Annual Meeting-1983.

Miller, R. B. Task taxonomy: Science or technology? *Ergonomics*1967,*10*,167-176.

Naylor, J. C., & Briggs, G. E. Effects of task complexity and task organization on the relative efficiency of part and whole training methods. *Journal of Experimental Psychology*, 1963, *65*, 217-224.

Seymour, W. D. Experiments on the acquisition of industrial skills. (Part 3) *Occupational Psychology*, 1956, *30*, 94-104.

Stammers, R., & Patrick, J. *The Psychology of Training*. London: Methuen, 1976.

Sternberg, S. On the discovery of processing stages: Some extensions of Donders' method. *Acta Psychologica*,1969,*30*,276-315.

Multiple Task Performance

N85
14560

UNCLAS

N85 14560 D25

REPRESENTING MULTIDIMENSIONAL SYSTEMS USING VISUAL DISPLAYS

Elizabeth J. Casey, Arthur F. Kramer and Christopher D. Wickens
Cognitive Psychophysiology Laboratory
University of Illinois
Champaign, Illinois 61820

SUMMARY

Techniques employed to represent multiattribute information in an integrated, object display are reviewed and discussed. A study is proposed to investigate the effects of system parameters such as inter-variable correlation on the choice of an optimal display format. The results of a psychophysical scaling study of five different displays are presented.

INTRODUCTION

A visual display acts as an interface between a dynamic system and a human operator. Its composition is critical to the performance of the operator in controlling a system and detecting and diagnosing system failures. As the complexity of systems has increased, the amount of information available to the human operator has become overwhelming. Therefore, there is a serious need to optimize the display formats used to present system status information. The operator must be presented with information in a format that requires a minimal amount of mental transformation prior to integrating it with an already existing internal model of the system. The display format should also allow the operator to respond quickly and accurately when so required.

When acting in a supervisory role, the human formulates a high fidelity internal model of the system. The internal model refers to the human operator's conception of the information structure and serves as a basis for potential actions (Wickens, 1984). A display compatible with the operator's internal model will minimize workload thus allowing faster, more accurate detection and diagnosis. The internal model may vary along several dimensions. These dimensions include the frequency with which the model is updated and the degree to which the representation of the system is spatial and/or verbal (Bainbridge, 1981; Landeweerd, 1979). Another dimension of variability is the perceived degree of integrality of the system variables, or in other words, the operator's perception of the relative correlations between the variables. In our research program we will examine several different methods of graphically representing a dynamic multiattribute system. We will attempt to match the display techniques so they are compatible with the operator's conception of the system.

One type of graphic representation of multivariate data which has recently received a great deal of attention is the object display in which several variables are typically represented on a single frame of reference. As an example, consider a polygon formed by connecting the ends of invisible lines which extend out from one point. The length of the imaginary spokes, and therefore the inner angles of the vertices of the polygon represent the values of the system attributes. In addition to giving information about the magnitude of each variable, the overall shape and size of this display can give insight into relationships between the variables. A practical application of the integrated presentation of multivariate data is found in the field of aviation. The contact analog display combines the two variables of roll and pitch into a single, highly schematic representation of the aircraft.

Some of the advantages of the object display over traditional, separate representations of multivariate systems include the subjects' familiarity with the objects, the holistic property of object perception by which subjects perceive the overall status of the system, and the single frame of reference against which all of the variables can be compared. We hypothesize that the integrated representation provided by the object display will aid the operator in perceiving the relationships among the system variables. This is because a lifetime's experience of dealing with objects and the correlated dimensions of these objects as they are transformed in space, has allowed us to associate the integral dimensions that define an object with a correlation between the values along those dimensions. We hypothesize that this association should allow better perception of correlated variables through integral displays. This hypothesis has received some validation in the earlier research of Garner (Garner, 1970; Garner & Fefoldy, 1970). Other research has shown that subjects are particularly sensitive to correlations between variables and thus a display which optimally depicts relational information will be useful to operators of complex, multidimensional systems (Medin, Alton, Edelson & Freko, 1982).

Several empirical studies have been conducted to assess the relative advantages and disadvantages of different displays. In one such study four displays were evaluated: arrays of digits, each digit defining a system variable; glyphs, which portrayed the system variables using the lengths of a series of rays surmounting a circle of fixed size; polygons, the distances from the center to the vertices representing the system variables; and schematic faces, in which each feature delineated a system variable (Jacob, Egeth & Bevan, 1976). Using a card sorting task and a paired associate learning task, Jacob et al. demonstrated that people process information from standard displays (such as the arrays of digits) in a "piecemeal, sequential mode which could obscure the recognition of relationships among the individual elements". In contrast, Jacob et. al. assert that the stimuli represented in object displays are processed holistically resulting in easier detection of relationships among variables.

In a series of studies conducted at the Idaho National Engineering Laboratory (INEL) investigators have evaluated the potential use of object displays as Safety Parameter Display Systems (SPDS) in nuclear power plant control rooms (Blackman et al., 1983; Danchak, 1981; Gertman et al., 1982; Petersen et al., 1982). The basic functions of the SPDS include; alerting the operator to the occurrence of abnormal plant conditions, aiding the operator in identifying specific abnormal parameters and assisting the operator in diagnosing plant conditions based on the relative values of parameters. The INEL studies, which have evaluated different object displays in a series of tasks and with several different methodological techniques (psychophysical scaling, multivariate rating scales, checklists and decision analysis), have shown that generally performance with object displays is equivalent or superior to that with more traditional, separate representations of multivariate data. Westinghouse has also proposed and evaluated an object display (polygon) as one of a series of displays to be used in an SPDS (Little & Woods, 1981).

A recent study demonstrated that using a polygon to display system information fostered subject performance that was superior to the performance obtained when the information was displayed on a bar graph (Carswell & Wickens, 1984). This effect was demonstrated under several different conditions. Object displays have also been found useful in presenting a multivariate frame of reference to identify relevant physiologic patterns that may delineate the seriousness of medical abnormalities (Siegel et al., 1971).

Several investigators have proposed that the holistic perception engendered by schematic faces would be ideal for the presentation of highly related system parameters (Danchak, 1981; Wilkinson, 1981). In one study concerned with the facial representation of multivariate data, the investigator found that the sterotype meaning already present in the faces could be measured and exploited to construct an inherently meaningful display (Jacob, 1978). Thus, in addition to the advantages already cited for object displays, subjects' familiarity with facial expressions appears to provide another dimension which can enhance the perception of multidimensional data. Schematic face displays have been found to be superior to separate numeric presentations of multivariate information in areas as diverse as the financial profile of businesses (Moriarity, 1979), Soviet foreign policy in Sub-Saharan Africa (Wang & Lake, 1978), the evaluation of psychiatric data (Mezzick & Worthington, 1978), and product performance (Hahn, Morgan & Lorensen, 1983).

Our program of research is concerned with explicating the factors that influence the subject's perception, transformation, and response to complex, visually presented information. We are pursuing this issue in the context of investigating the conditions under which different displays provide an optimal representation of system status information. Some of the specific issues which we will address include:

1) Are displays which provide an integrated representation of system parameters superior to more traditional displays which present the same information separately (i.e. polygons vs. meters)? Furthermore, does the display format interact with the type of task which the operator is required to perform? Some research has suggested that polygons may be superior to meters for detection tasks while meters appear to be optimal for the localization of abnormal variables (Petersen et al., 1981, 1982).

2) Does the correlational structure of the system variables interact with the presentation format of the variables? In other words, are different display formats optimal for systems with different inter-variable correlations? Highly integrated object displays have been proposed to be most useful in situations in which the system parameters are moderately to highly correlated (Wickens, 1984).

3) Do subjects with different degrees of spatial ability adopt different strategies to perform detection and diagnosis tasks? Can we optimize the subjects' performance by presenting system status information in a manner consistent with the subjects' preferred processing strategy?

In our program of investigation, five different displays will be evaluated. These displays, which are represented on a continuum of integrality of information presentation, are shown in Figure 1. The displays include digital meters, bar graphs, glyphs, polygons (pentagons), and schematic faces. Each of the displays will represent five variables of a dynamically changing system. The five facial features that will represent the system parameters are the angle of the eyebrows, the width of the mouth, and the lengths of the eyes, the ears, and the nose.

In a series of experiments, we intend to investigate the relative merits of these different displays in tasks which involve monitoring a dynamic, multivariate system. The tasks will include detection and diagnosis of system failures. While we hypothesize that face and object displays will foster better performance to the extent that variables are correlated, it is essential that certain basic factors of the displays be accounted for before such studies are run.

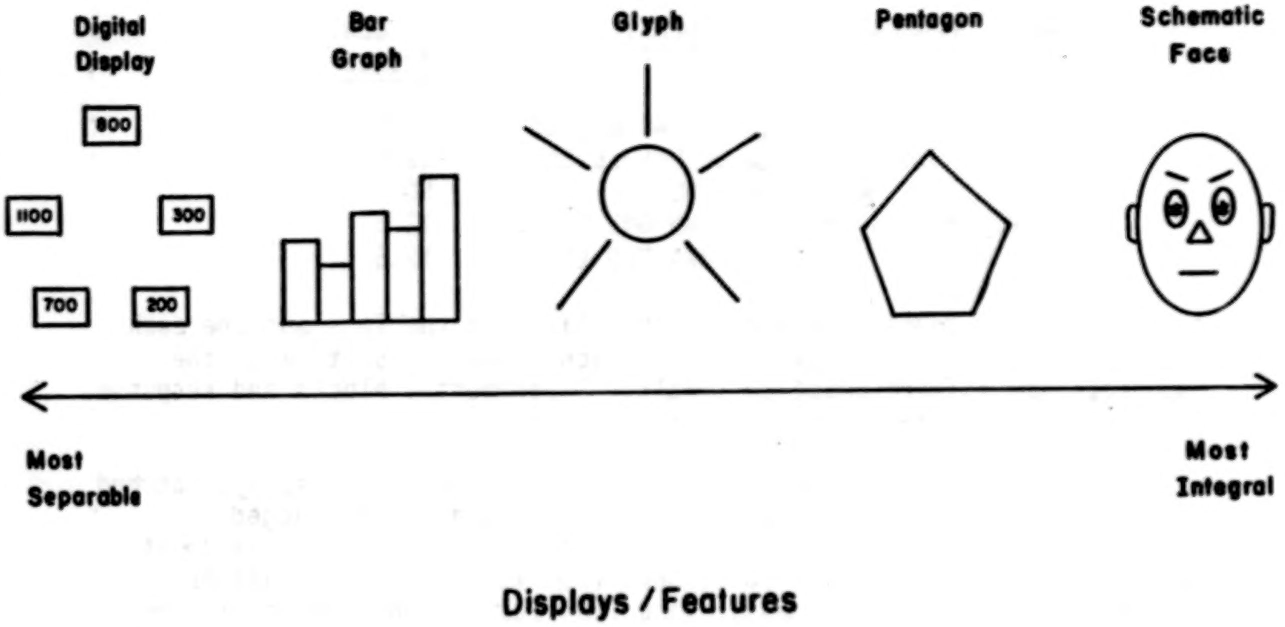

Digital Display
800
1100
300
700
200
Bar Graph
Glyph
Pentagon
Schematic Face
Most Separable
Most Integral
Displays / Features

Figure 1

Scaling Study

It is important in any comparison of visual displays to determine whether the superiority of a given visual display can be accounted for by perceptual factors. Therefore, in order to properly compare these displays we first had to ensure that it was equally difficult to perceive a change in a system variable regardless of the display or display feature on which that variable was represented. Thus, a psychophysical scaling study was conducted on the five displays.

Ten college students participated in the experiment. All were right-handed with normal or corrected-to-normal vision. During the two hour session the subjects performed the task with each of the five displays. The subject's task was to decide whether two sequentially presented displays matched or mismatched. The importance of both speed and accuracy was emphasized. Subjects pressed one response button if the displays matched and another if they did not. In a single block of trials only one of the five variables on a display was to be attended by the subject. The other four variables remained at constant levels. Each of the five features of the face varied in different blocks.

For each display and feature, ten equidistant levels were defined. The "standard" (S1) was always either at level 5 or at level 11. The comparison stimulus (S2) varied as follows:

STANDARD	PERCENT TRIALS	COMPARISON	PERCENT TRIALS
		UP TWO LEVELS	12.5
LEVEL 5	50.0	UP ONE LEVEL	12.5
OR		SAME	50.0
LEVEL 11	50.0	DOWN ONE LEVEL	12.5
		DOWN TWO LEVELS	12.5

In total there were nine blocks, five for the face and one each for the other four displays. Before each block of 160 trials, the subjects had fifteen practice trials. Experimental blocks and response buttons were counterbalanced across subjects.

The amount of time required to decide whether two displays matched or mismatched was affected by the type of display being judged (F(8,72)=4.3, p<.01). The order of displays from fastest to slowest was the meters (318 msec), bar graphs (334 msec), polygon (342 msec), glyphs (373 msec) and schematic face (394 msec). The amount of time required to compare different facial features ranged from 375 msec for the eyebrows to 417 msec for the mouth. RT was also influenced by stimulus level. Slight mismatches took longer to respond to than matches and more obvious mismatches (F(4,36)=18.3, p<.01). There was an interaction between display type and stimulus level such that RT performance with meters and bar graphs was not differentially affected by stimulus level (F(32, 288)=1.7, p<.01). Error rate generally followed the same pattern as RT with larger error rates being associated with longer RTs.

The results of this scaling study will be used to adjust the magnitudes of the physical changes of the display components. The ranges of variations will be scaled to be psychophysically equivalent. Thus, any differences in performance among displays will not be attributed to perceptual factors.

ACKNOWLEDGEMENTS

This research is supported by contract #MDA903-83-K-0255 from the Army Research Institute with Dr. Marshall Narva as technical monitor. We gratefully acknowledge the programming assistance of Mark Klein.

REFERENCES

Bainbridge, L. (1981). Mathematical equations or processing routines? In J. Rasmussen and W.B. Rouse (Eds.), Human Detection and Diagnosis of System Failures. New York: Plenum Pres.

Blackman, H.S., Gertman, D.I., Gilmore, W.E. and Ford, R.E. (1983). Noninteractive simulator evaluation for CRT-generated displays. Idaho National Engineering Laboratory. NUREG/CR-3556.

Carswell, C.M. and Wickens C.D. (1984). Stimulus integrality in displays of system input-output relationships: A failure detection study. In E. Alluisi (ed.) *Proceedings. 28th Annual Meeting of the Human Factors Society*, Santa Monica, CA: Human Factors Press, 1984.

Danchak, M.M. (1981). Techniques for displaying data on cathode ray tubes with applications to nuclear process control. Idaho National Engineering Laboratory. NUREG/CR-1944.

Garner, W.R. (1970). The stimulus in information processing. *American Psychologist*, *25*, 350-358.

Garner, W.R., & Fefoldy, G.L. (1970). Integrality of stimulus dimension in various types of information processing. *Cognitive Psychology*, *1*, 225-241.

Gertman, D.I., Beckman, H.S., Banks, W.W. and Petersen, R.J. (1982). CRT display evaluation: The multidimensional rating of CRT generated displays. Idaho National Engineering Laboratory. NUREG/CR-2942.

Hahn, G.J., Morgan, C.B. and Lorensen, A.E. (1983). Color face plots for displays of product performance. *IEEE Computer Graphics and Applications*, *3*, 23-29.

Jacob, J.K. (1978). Facial representation of multivariate data. In P.C.C. Wang (Ed.), *Graphical Representation of Multivariate Data*. New York: Academic Press.

Jacob, J.K., Egeth, H.E., and Bevan, W. (1976) The Face as a Data Display. *Human Factors*, *18*, 189-200.

Landeweerd, J.A. (1979). Internal representation of a process, fault diagnosis and fault correction. *Ergonomics*, *22*, 1343-1351.

Little, J.L. and Woods, D.D. (1981). A design methodology for the man-machine interface for nuclear power plant emergency response facilities. *Proceedings of the Joint American Nuclear Society and Taiwan Nuclear Society Conference on Human Factors and Safety*. Taipei, Taiwan.

Medin, D.L., Altom, M.W., Edelson, S.M., and Freko, D. (1982). Correlated Symptoms and Simulated Medical Classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 37-50.

Mezzick, J.E. and Worthington, D.R.L. (1978). A comparison of graphical representations of multivariate psychiatric diagnostic data. In P.C.C. Wang (Ed.), Graphical Representation of Multivariate Data. New York: Academic Press.

Moriatrity, S. (1979). Communicating finanical information through multidimensional graphics. Journal of Accounting Research, 17, 205-224.

Petersen, R.J., Banks, W.W. and Gertman, D.I. (1981). Performance based evaluation of graphic displays for nuclear power plant control rooms. Proceedings of the Conference on Human Factors in Computer Systems. Gaithersburg, Maryland.

Petersen, R.J., Smith, R.L., Banks, W.W. and Gertman, D.I. (1982). An empirical examination of evaluation methods for computer generated displays: Psychophysics. Idaho National Engineering Laboratory. NUREG/CR-2942.

Siegel, J.H., Goldwyn, R.M. and Freidman, H.P. (1971). Pattern and process in the evolution of septic shock. Surgery, 70, 232-245.

Wang, P.C.C. and Lake, G.E. (1978). Application of graphical multivariate techniques in policy sciences. In P.C.C. Wang (Ed.), Graphical Representation of Multivariate Data. New York: Academic Press.

Wickens, C.D. (1984). Engineering Psychology and Human Performance. Columbus, Ohio: Charles E. Merrill Publishing Company.

Wilkinson, L. (1981). An experimental evaluation of multivariate graphical point representations. Proceedings of the Conference on Human Factors in Computer Systems. Gaithersburg, Maryland.

Woods, D., Wise, J. and Hanes, L. (1981). An evaluation of nuclear power plant safety parameter display systems. In R.C. Sugarman (Ed.), Proceedings of the 25th Annual Meeting of the Human Factors Society. Santa Monica, California.

N85
14561

UNCLAS

N85 14561 D26

TYPES OF TRACKING ERRORS INDUCED BY CONCURRENT SECONDARY MANUAL TASK

Stuart T. Klapp, Patricia A. Kelly,
Vernol Battiste, and Sherry Dunbar

California State University, Hayward
Hayward, CA. 94542

Future one-man helicopters may require the pilot to control flight with one hand, and simultaneously manipulate other instruments using the other hand. This report of work in progress examines the nature of errors induced in a right hand tracking task (simulating flight control) when responses are required by the left hand. The present experiment focused on detection of hesitations in which the tracking joy stick remained motionless for 1/3 sec. or longer.

METHOD

The 12 subjects were right handed students who participated as one option of a course requirement. The right hand task was pursuit tracking in which the position of the joy stick corresponded to the vertical position of a cursor on a CRT display. The subject was to attempt to keep this cursor within a target box which moved up and down, driven by a random-appearing forcing function which changed velocity no more often than once per 167 msec. At randomly determined times (average rate of once per 30 sec.) a tone sounded signaling that a lever-moving response should be executed by the left hand.

Tracking performance was observed on corresponding forcing function segments with and without the left hand response. In order to detect hesitations, the position of the tracking joy stick was sampled 60 times per sec. during the first 2 sec. after a stimulus for a left hand response, and during the corresponding control right-hand only segments.

The left hand responses were movements of a switch handle at least 1.0 mm. to the left or right as commanded by a tone. This tone was terminated when the correct response was completed. A high-frequency tone (2000 Hz.) signaled a rightward movement and a low-frequency tone (500 Hz.) signaled a leftward movement. The timing of these signals was random for all subjects. For half of the subjects the appearance of high and low tones was also random, and hence the left hand task was a choice RT situation. For the remaining subjects the high and low tones appeared in predictable blocks, so that the left hand task was a simple RT situation.

N85 14561

For 1/3 of the subjects, the tracking task was emphasized by instruction, and by the presence of an unpleasant auditory alarm which sounded when the cursor was beyond the boundaries of the target box. For another 1/3 of the subjects the left hand response was emphasized by instruction, and by the same alarm which sounded when the incorrect directional response was made. The remaining 1/3 of the subjects received no emphasis instruction and no alarm.

RESULTS

Tracking (right hand).

We were primarily interested in the occurrence of hesitations in tracking. A hesitation was defined as holding the tracking joy stick motionless as determined by sampling its position at the rate of 60 observations per sec. In order to qualify, a hesitation had to start no later than 1 sec. after the stimulus for the left hand response (or within the corresponding time in control segments) and to last for 1/3 sec. or more. Hesitations, or portions of hesitations, which resulted in minimal difference between cursor position and center of the target (less than 1/2 of the upward or downward tolerance as defined by target box height) were considered to be correct responses rather than error hesitations.

Corresponding to each opportunity for the right hand to hesitate as a result of the concurrent left hand response, there was a control tracking pattern in which the same forcing function segment was tracked without a stimulus for left hand responding. Hesitations occured on 48% of the opportunites when the left hand stimulus was present, but on only 6.5% of the control opportunities, F(1,11)=27.0, p <.001. Clearly hesitations were caused by the left hand responses, and were frequent rather than rare events.

The remainder of this discussion concerns those hesitations which occurred on the dual task trials. Table 1 displays the distribution of the durations of these hesitations. (By definition only hesitations exceeding 333 msec. are considered.)

Duration range, msec.	Observations
333-480	33
481-666	28
667-833	17
834-1000	5
> 1000	5

Table 1. Distribution of hesitation durations.

The emphasis instruction and alarm influenced the rate of hesitation in right hand tracking. Hesitations occured on 76% of the opportunities when left hand performance was emphasized, but only on 29% of the opportunities when right hand tracking was emphasized, F(1,6)=7.4, p <.05. An intermediate rate of hesitations, 37%, occurred with no emphasis instruction or alarm. The two subjects with fewest hesitations were both in the track emphasis condition, and the two subjects with the most hesitations were both in the left hand emphasis condition.

We had thought that hesitations might by attributable to the necessity of making decisions concerning which response to generate with the left hand. Thus, we expected to find more right hand hesitations for the choice RT left hand condition than for the simple RT left hand condition. The only other study we know of reporting hesitations (Cliff, 1973) observed hesitations upon inspection of tracking records when subjects engaged in speech shadowing, a task for which the simple-choice distinction is not clear. Contrary to our expectations, there was no hint of reduced hesitations for simple RT left hand responses. The rate of hesitations was 53% for simple compared to 42% for choice.

Left hand performance.

For the left hand response, error rate (initially incorrect direction of movement) and median RT was determined for each subject in each condition. Consistent with the usual finding, overall (mean of medians) RT was longer for the choice RT condition (799 msec., error rate 10.5%) than for simple RT (491 msec., error rate 2.3%), F(1,10)= 12.9, p<.005. For the simple RT task, subjects for whom the left hand task was emphasized produced a shorter mean RT (378 msec.) compared to the other emphasis conditions (554 for track emphasis and 542 msec. for no empahsis). By contrast, for the choice RT task, emphasis of the left hand task produced a longer RT (991 msec.) compared to the other emphasis conditions (625 and 670 msec.). This apparent interaction may be understood by assuming that, for choice RT and alarmed left hand, subjects confirmed their response selection before moving the response switch, thereby increasing RT. By contrast, for simple RT, no response selection was required and emphasis on the left hand response caused a reduction in RT. We report these RT results primarily to indicate that RT behaved in an orderly and understandable manner for these routine comparisons. This sets the stage for our primary point about RT for which the results were rather unexpected.

Data on left hand performance was separated into two pools, one for those trials on which the right hand tracking task exhibited a hesitation and one for those trials on which no hesitation occurred. Surprisingly, left hand performance did not differ significantly as a function of this distinction. The mean left hand RT was 629 msec. (error rate 4.2%) when the right hand hesitated, compared to 648 msec. (error rate 4.6%) with no hesitation. These RT means did not differ significantly, F(1,11)<1. Thus, there is little reason to suppose that subjects trade off between the left and right hand performance. The other report of tracking hesitations (Cliff, 1973) also indicated that secondary task performance did not differ as a function of whether tracking showed a hesitation.

DISCUSSION

Hesitations in flight control would represent a potential disaster in nap of the earth flight. Thus, the occurrence of tracking hesitations of at least 1/3 sec. duration on nearly half of the instances in which a concurrent left hand response was required is a cause of concern. Although hesitations occurred less often when emphasis was placed on tracking, our emphasis instruction and alarm were not successful in eliminating hesitations completely. Thus, it is possible that pilots might hesitate in flight control even in dangerous situations.

Contrary to our expectation, the rate of tracking hesitations was independent of whether the concurrent left hand response was a choice RT or simple RT task. Thus, the hesitations were not produced by response selection. Another unexpected result was that hesitations on tracking did not produce improvement in left hand performance. Therefore, we propose a model which does not assume that hesitations result from diverting a portion of limited resources to left hand response selection.

Our model assumes a single channel of attention which can be directed toward the right hand response or toward the left hand response, but which cannot simultaneously attend to both responses. Diversion of attention to the left hand occurs on all trials which require a left response, including those for which the right hand does not hesitate. This attention shift occurs even if left hand response selection is predetermined (simple RT task). When attention is diverted to the left hand, right hand tracking continues in an open loop mode of control until completion of a response segment which has been "programmed" (Keele, 1968; Klapp, 1975). If the program in "buffer memory" (Klapp,1981) is completed prior to return of attention to the right hand, a hesitation occurs.

This much of the model accounts for several observations. First, in a separate pilot study of repetitive tapping with the right hand, we observed open loop continuations rather than hesitations when a concurrent left hand response was required. This finding can be understood by assuming that the repetitive nature of the tapping response corresponded to a long-duration response program for open loop continuation. This hypothesis also accounts for the observation that no hesitations in the tracking task started until about 350 msec. after the stimulus for left hand responding. This delay before hesitation corresponds to the lag of the tracking response behind the stimulus. If this lag represents the minimum duration of preprogrammed tracking, then, according to the model, no hesitations should occur until this programmed response sequence is completed.

In this model, attention is assumed to be completely diverted to the left hand on all trials involving the left hand. Thus, the model accounts for the result that left hand performance was independent of whether the right hand exhibited a hesitation. Hesitations occur if the right hand response program is too short, and the duration of the program is determined prior to allocation of attention to the left hand.

How, then, are we to account for the apparently contradictory finding that, although left hand performance was independent of whether a particular trial included a hesitiation, nevertheless subjects responded to the emphasis instruction and alarm by reducing hesitations? To handle this result the model is elaborated by assuming that subjects can reprogram the right hand tracking response just before they divert attention away from tracking. This assures that a long program is available for the right hand, thereby reducing the chance that the program will be completed prior to return of attention to the right hand. This reduces hesitations. But, programming takes time (Klapp, 1975,1981) and hence delays the left hand response. Note that this view accounts for overall sensitivity to the emphasis manipulation without incorrectly predicting a trade-off between hesitations and left hand performance on a trial by trail basis within an emphasis condition.

This model has features which are optimistic and others which are pessimistic concerning the possibility of maintaining flight control in a dual task situation. On the optimistic side is the possibility that, on a task with preview (such as flight control), programming in advance might be rather extensive and hence hesitations might be greatly reduced if pilots were made aware of the necessity of planning a flight control sequence before attempting a left hand secondary response. On the pessimistic side, this model is a single channel view which holds that flight control can only operate open loop during performance of the left hand response. Hence it would not be possible to respond to unexpected flight events while the secondary task is being performed.

REFERENCES

Cliff, R.C. (1973) Attention sharing in the performance of a dynamic dual task. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, 241-248.

Keele, S.W. (1968) Movement control in skilled motor performance. *Psychological Bulletin*, 70, 387-403.

Klapp, S.T. (1975) Feedback versus motor programming in the control of aimed movements. *Journal of Experimental Psychology: Human Perception and Performance*, 1, 147-153.

Klapp, S.T. (1981), Motor programming is not the only process which can influence RT: Some thoughts on the Marteniuk and MacKenzie analysis. *Journal of Motor Behavior*, 13, 320-328.

ACKNOWLEDGEMENT

This research was sponsored by NASA-Ames Cooperative Agreement NCC 2-223. Dr. E. James Hartzell was technical monitor, and his advice and insights are acknowledged with appreciation. We also thank George Eggleton, Dr. Christopher Morgan, and John Tyler for the design and constuction of the specialized computer system used to collect the data.

N85
14562

UNCLAS

N85 14562

D27

THE EFFECTS OF TASK STRUCTURES ON TIME-SHARING EFFICIENCY AND RESOURCE ALLOCATION OPTIMALITY

Pamela S. Tsang*
Department of Psychology
University of Illinois at Urbana-Champaign
Champaign, Illinois

Christopher D. Wickens
Institute of Aviation and Department of Psychology
University of Illinois at Urbana-Champaign
Willard Airport
Savoy, Illinois

ABSTRACT

A distinction was made between two aspects of time-sharing performance: time-sharing efficiency and attention allocation optimality. The first is concerned with the level of joint performance of the time-shared tasks. The second is concerned with the consistency of protecting the performance of a high priority task from varying with changes in task demand. A secondary task technique was employed to evaluate the effects of the task structures of the component time-shared tasks on both aspects of the time-sharing performance. Five pairs of dual tasks differing in their structural configurations were investigated. The primary task was a visual/manual tracking task which requires spatial processing. The secondary task was either another tracking task or a verbal memory task with one of four different input/output configurations. Congruent to a common finding, time-sharing efficiency was observed to decrease with an increasing overlap of resources utilized by the time-shared tasks. Results also tend to support the hypothesis that resource allocation is more optimal when the time-shared tasks placed heavy demands on common processing resources than when they utilized separate resources. These data suggest that careful consideration of the tradeoff between time-sharing efficiency and resource allocation optimality is necessary in making multitask design decisions.

INTRODUCTION

Imagine a scenario in which the pilot is tracking a landing beam, simultaneously communicating with the ground controller, and all the while, coping with unseen gusts. In such dynamic high workload situations, what might the effects of the task structures of the many concurrent tasks on the pilot's time-sharing performance be? Two predictions are offered by the structure-specific resource model

---

* Now a National Research Council Research Associate at NASA-Ames Research Center, Moffett Field, California.

(Wickens, 1980). In this model, task structures are defined by three dichotomous dimensions related to: (a) the stages of processing (perceptual/central vs. response processing), (b) the codes of processing (spatial vs. verbal processing), and (c) the input/output (I/O) modalities (visual vs. auditory/manual vs. speech). Each of these elements is postulated to be associated with a separate pool of resource and the degree of resource overlap between two time-shared tasks is defined by the number of task structures they have in common.

The structure-specific resource model predicts that a greater time-sharing efficiency can be achieved when the time-shared tasks are structurally different and place heavy demand on different resources than when they are structurally similar and have to compete for the same resources. There are two reasons for this. First, there are potentially more resources available in the separate resource case. Second, because the different resources are conceptualized as rather independent pools of attentional capacity, less task interference is expected between separate resources than within a common resource. On the other hand, the model also predicts that continuous resource allocation would only be possible if the time-shared tasks place heavy demand on at least some common resources. This is because, according to the model, separate resources are not sharable and what is withdrawn from one type of resource should not benefit the other. Although both predictions have already received some empirical support (e.g., Brickner & Gopher, 1981; Triesman & Davies, 1973; Wickens, Tsang, & Benel, 1979), there is yet little research effort to systematically examine the effects of the task structures on the relation or interaction between the two aspects of time-sharing performance.

In the present paper, time-sharing efficiency and attention or resource allocation optimality are considered to be two equally important aspects of the more generic term, time-sharing performance. However, a major point here is how they can be and why they should be distinguished. The latter is to be demonstrated by the experiment described below and the former is a matter of definition. Time-sharing efficiency describes the maximum joint performance of the time-shared tasks whereas resource allocation optimality describes the control of the amount of resources to be distributed among the time-shared tasks. Operationally, time-sharing efficiency is measured by the degree of task interference between the time-shared tasks whereas resource allocation optimality is inferred from the consistency of maintaining the performance at a desired level regardless of demand fluctuations. In some instances (e.g., under very high workload situations), optimal allocation may mean maximizing one performance while sacrificing the other. In contrast, time-sharing efficiency is always concerned with the least performance decrement to both tasks as a whole. It is hypothesized in the present paper that the level of allocation optimality will increase, but the degree of time-sharing efficiency will decrease, with an increasing degree of resource overlap between the time-shared tasks. This hypothesis will be tested by contrasting the level of allocation optimality and time-sharing efficiency attainable in five pairs of dual tasks differing in their degree of similarity in task structures between the component time-shared tasks.

EXPERIMENTAL APPROACH

A secondary task technique in which a high priority primary task is time-shared with a low priority secondary task will be employed. With this technique, the primary task performance is to be kept constant at the same level as its single task performance (see Ogden, Levine, & Eisner, 1979) so that the differential secondary task performances between the single and dual task conditions can be used as an index of the demand imposed by the primary task.

All five pairs of dual tasks will have a compensatory tracking task as the primary task whose difficulty (control order) varies dynamically within a trial. One pair has a constant difficulty compensatory tracking task as the secondary task. The secondary task of the other four pairs has a constant difficulty running memory task in all four possible combinations of two input (visual and auditory) and two output (manual and speech) modalities. The tracking task is considered to impose the greatest demands upon the response stage but also requires increased perceptual/central processing resources of a spatial nature with higher order control dynamics (see Israel, Chesney, Wickens, & Donchin, 1980). In contrast, the running memory task is assumed to place heavy demands on the central processing stage and relatively less on the response resources. Because of the type of stimuli used, it is considered a verbal task. Therefore, the tracking task and the memory task differ from each other in terms of both the processing stages (response vs. perceptual/central) and processing codes (spatial vs. verbal). Furthermore, the four memory-tracking pairs differ from each other in their various combinations of the I/O modalities.

Placing these five pairs of dual tasks on a continuum of degree of overlapping resources between the time-shared tasks, the dual tracking task (being a pair having identical task structures) will be on one extreme and the auditory/speech memory-tracking pair (being a pair with two tasks relying on different processing stages and codes, and completely separate I/O modalities) will be on the other extreme. This continuum is portrayed in Figure 1 with the different secondary tasks labelling the horizontal axis. As shown on this figure, time-sharing efficiency is expected to increase as the degree of resource overlap decreases. In contrast, allocation optimality is expected to increase in the opposite direction.

The relative resource demands of the time-shared tasks was manipulated by: (1) changing the task priorities by means of payoffs and instructions, and (2) varying the difficulty level of the primary task (for example see, Gopher & Navon, 1980; Kantowitz & Knight, 1976). The underlying rationale for these manipulations is as follow. As the priority or the difficulty level of the primary task increases, additional resources will have to be invested in the primary task in order to maintain its performance at a constant level. In the situation where the primary and secondary tasks must compete for the same

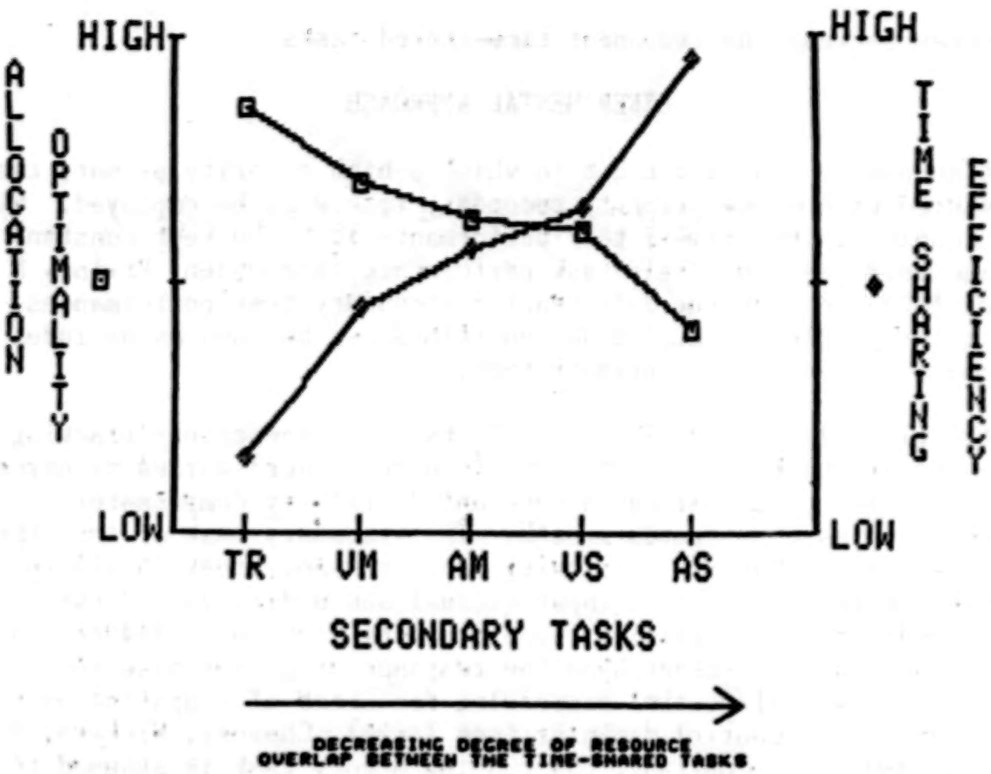


Figure 1. Predicted time-sharing efficiency and resource allocation optimality obtained from five pairs of dual tasks with various degrees of structural similarity between the component time-shared tasks. All five pairs had a visual/manual tracking task as the primary task.

resources, the secondary task performance will inevitably deteriorate because of its decreased share of resources, provided that the maximum available capacity is already being deployed.

A departure from the more conventional practice of employing a discrete manipulation of the relative use of resources between trials is the adoption of a continuous difficulty manipulation within a trial (time-varying primary task difficulty) in the present study to reflect a more realistic dynamic environment (see also Wickens & Tsang, 1979). Since the primary task performance is to be kept constant, the primary task difficulty momentary increase is thus to be absorbed by the secondary task. The secondary task error is therefore expected to covary more closely with the primary task difficulty than the primary task error would if the two tasks were competing for the same resources. Hence, with perfect control of resource allocation, there should be little correlation between the primary task performance and its difficulty variation but a relatively high correlation between the secondary task performance and the primary task difficulty fluctuations. The correlations between the difficulty fluctuation and the primary and secondary tracking error will be estimated by the coherence measures obtained between the primary task difficulty function and the moment by moment tracking error through a bivariate time-series analysis (see Pierce & Wickens, 1978).

METHOD

Ten right-handed male subjects participated in the experiment. The tracking difficulty parameter was the percent of the second order component in a linear combination of first and second order of control dynamics. The primary task difficulty was always time-varying with the difficulty function varying between 0 (first order) and 1 (second order) at two constant rates within a trial (200 seconds). The secondary tracking difficulty was fixed at .5 for the entire trial. The secondary memory task was similar to the one employed by Zeitlin and Finkelman (1975). Digits from 0 to 9 were presented in a random order one at a time throughout the trial. Subjects were to recall the digits one-back as soon as the next stimulus appeared. The digits were presented either visually (V) on the same CRT display as the tracking task or auditorily (A) through headphones. Subjects responded either by saying the digits (speech response, S) or by pressing the appropriate button on a keyboard (manual response, M). The displays and response processing were controlled by a PDP 11/40 computer.

Subjects received four hours of single task practice which included training with the use of the speech recognition unit (Centigram Corporation Mike-2). Subjects then received three hours of dual task practice with all five dual tasks: tracking-tracking (TR-TR) and four pairs of memory-tracking (VM-TR, AM-TR, VS-TR, and AS-TR). Throughout training, subjects were simply asked to keep their error of both tasks as low as possible. Subjects were also encouraged to respond as quickly as possible without sacrificing accuracy for the memory task. Starting from session 8 and for the following three 1-hour sessions, each subject was instructed to maintain the primary task performance at the same level as his own best single task performance (which served as his own performance standard). Instructions and monetary payoffs emphasized that the most important objective was to maintain the primary task performance constant at the standard level. The secondary objective (with a much smaller monetary incentive) was to maximize the secondary task performance. The latter provision was included to discourage subjects from neglecting their secondary task entirely except when necessary to do so to protect the primary task.

Tracking performance measures included root mean square error (RMSE) and linear coherence measures (between the tracking error and primary task difficulty variation) derived from time-series analysis. Performance measures for the memory task included recall accuracy (percent error), and reaction times (RT).

RESULTS

Effects of Task Structures on Time-sharing Efficiency

The effects of the priority instructions on the dual tracking performance are portrayed in the average RMSE plotted in Figure 2. Before the priority instructions were introduced (Sessions 6 and 7), the levels of the primary and secondary task errors were almost

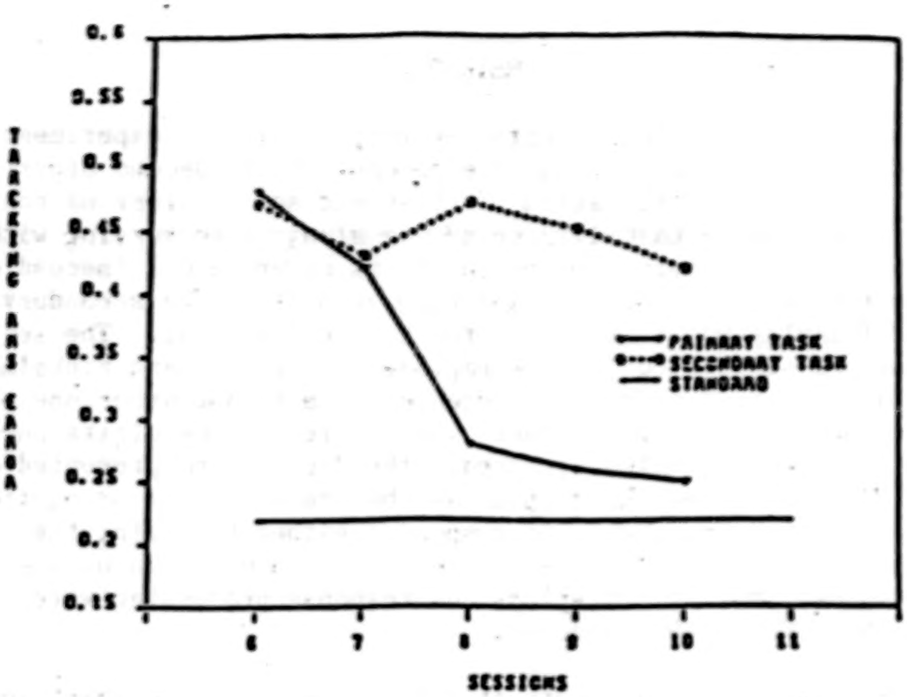


Figure 2. Dual tracking performance.

indiscriminable. With the introduction of the priority instructions in Session 8, the primary task error dropped drastically and remained fairly close to the average standard (with a RMSE of .22 as indicated by the horizontal line on the figure). In contrast, the secondary task error even increased slightly with the introduction of the priority instructions (from Session 7 to Session 8). Although the secondary task error eventually decreased as well, the decrease was much smaller compared to that of the primary task. The Task (primary vs. secondary) x Session interaction was significant at .01 level (_F_ (4,36) = 18.97). This interaction suggests that the subjects were voluntarily allocating more resources to the primary task than to the secondary task in response to the priority instructions. Furthermore, the continual, though slight, decrease in the secondary task error suggests that, with practice, only sufficient resources were allocated to the primary task to maintain its performance at the standard level; the spare resources were utilized to improve the secondary task performance.

The primary task performance of the four memory-tracking pairs also showed remarkable improvement with the priority instructions and they are compared with that of the dual tracking condition in Figure 3. Results of a two-way ANOVA (Session x Task) indicates that the session main effect (_F_ (4, 36) = 104.76) and the task main effect (_F_ (4, 36) = 20.01) were significant at .01 level. A post hoc pairwise comparisons (Scheffe) performed on the data from the last three sessions showed that the TR, VM, and AM conditions were not significantly different from each other ($p > .05$). However, these three conditions, all of which employed

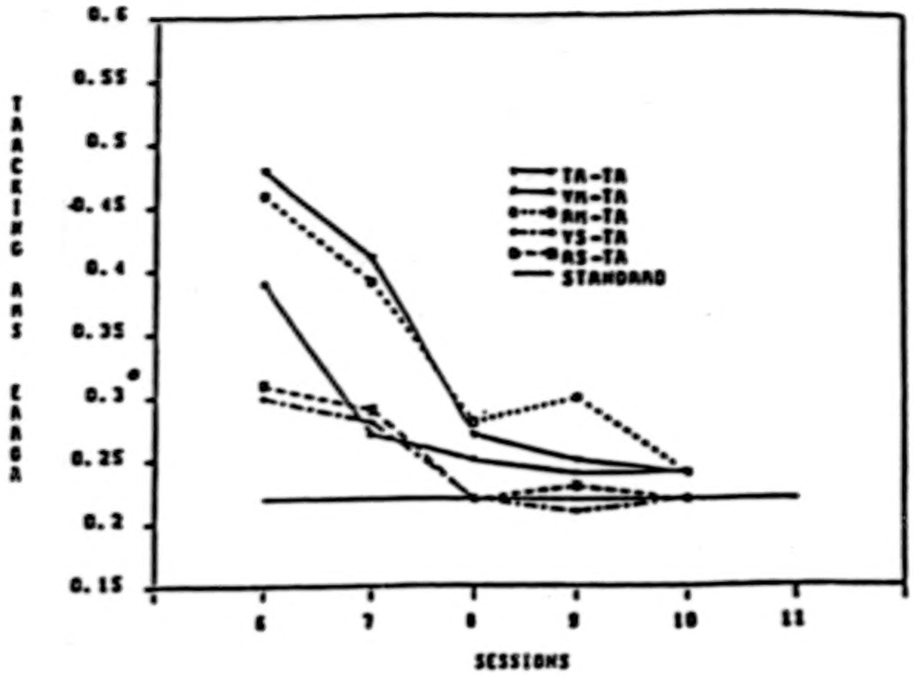


Figure 3. Primary task tracking performance of five structurally different task pairs.

a manual secondary response, had significantly higher error than the two conditions employing a speech secondary response (VS and AS) at .05 level.

Like the dual tracking condition, subjects were able to maintain a constant primary task performance at the standard level only at the expense of the secondary task for the memory-tracking task pairs. Reaction time decrements of the memory tasks are shown in Figure 4. The decrements were much reduced with practice, but they were not eliminated entirely even for the AS-TR pair. Although the task main effect was not significant ($p > .05$), Figures 3 and 4 together show that the degree of performance decrement (i.e., the extent of task interference) for each of the I/O conditions appears to be in the exact order predicted by the structure-specific resource model (see Figure 1) by the end of the experiment: The VM task had the greatest degree of shared resources with the tracking task and the greatest degree of task interference was observed between this task pair; the AS task had the least common resources with the tracking task and its RT decrement was the smallest; the AM and VS tasks each had one I/O modality in common with the tracking task and their RT decrements were found to be in between the two VM and AS extreme conditions.

Only the RT performance was analyzed because, with the exception of the first two hours of single task training, the percent error for most subjects was close to zero throughout the experiment.

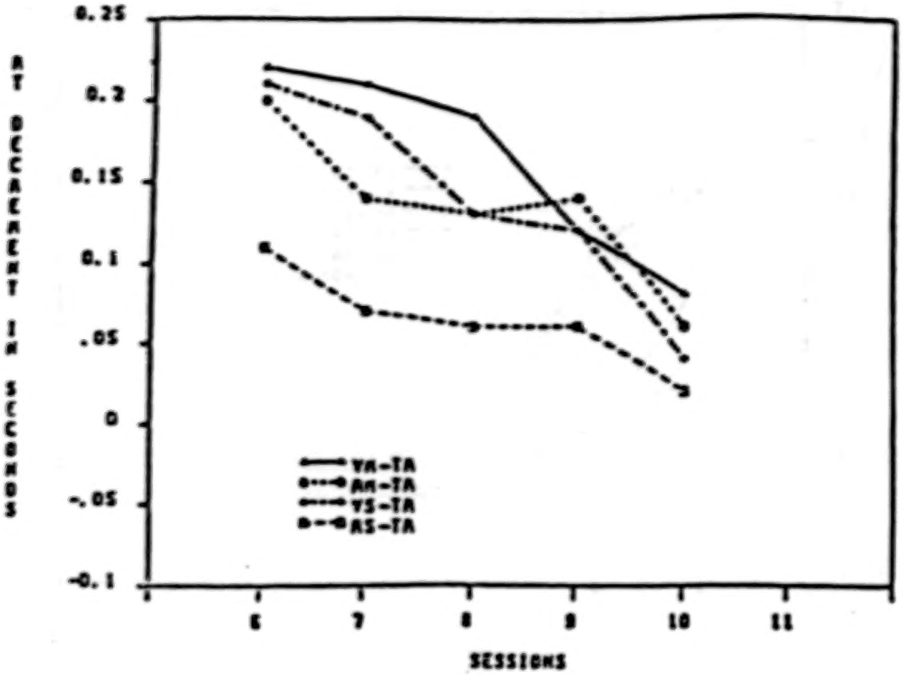


Figure 4. Reaction time decrements in the four memory-tracking conditions.

Effects of Task Structures on Resource Allocation Optimality

The raw sampled RMSE values (one per 50 ms) were smoothed by computing a running average of these values within a 2-second sliding window. The 200 averages were computed for every trial. These averages were ensembled across subjects before they were entered into the time-series analysis (BIOMED-02T). Figure 5 is a sample of such ensemble averages obtained under the AS-TR condition when the priority instructions were first introduced (Session 8). Coherence measures were obtained between the primary task difficulty variation and its error fluctuations for all five pairs of dual task for each session. Coherence measures between the primary task difficulty variation and the secondary task error were also obtained for the dual tracking conditions. However, time-series analysis was not performed on the RT data because of its discrete nature.

The dual tracking condition will be discussed first. As shown in Figure 6, an increase in the secondary task coherence measures was observed upon the introduction of the priority instructions in Session 8. This would indicate an increase in the secondary task performance variability that was time-locked to the primary task difficulty in spite of the fact that the secondary task difficulty was fixed at a constant level. Although the primary task coherence measures did not decrease in Session 8 as predicted by the optimal allocation model (Pierce & Wickens, 1978), its increase was slight compared to that of the secondary task. Unfortunately, since all the coherence measures were

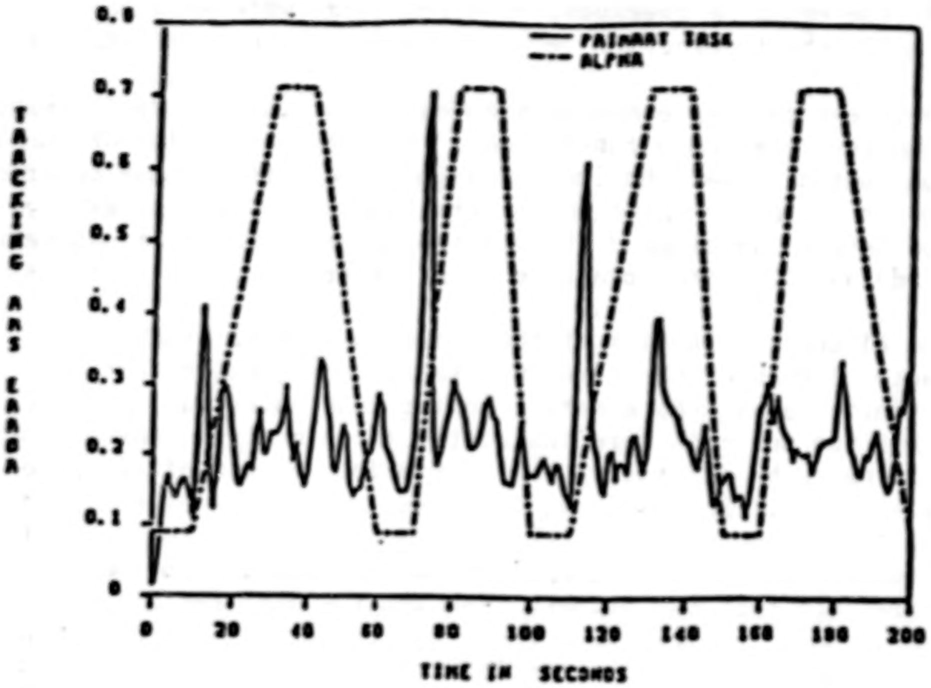


Figure 5. Primary task ensemble averages of the AS-TR condition obtained when the priority instructions were first introduced (Session 8).

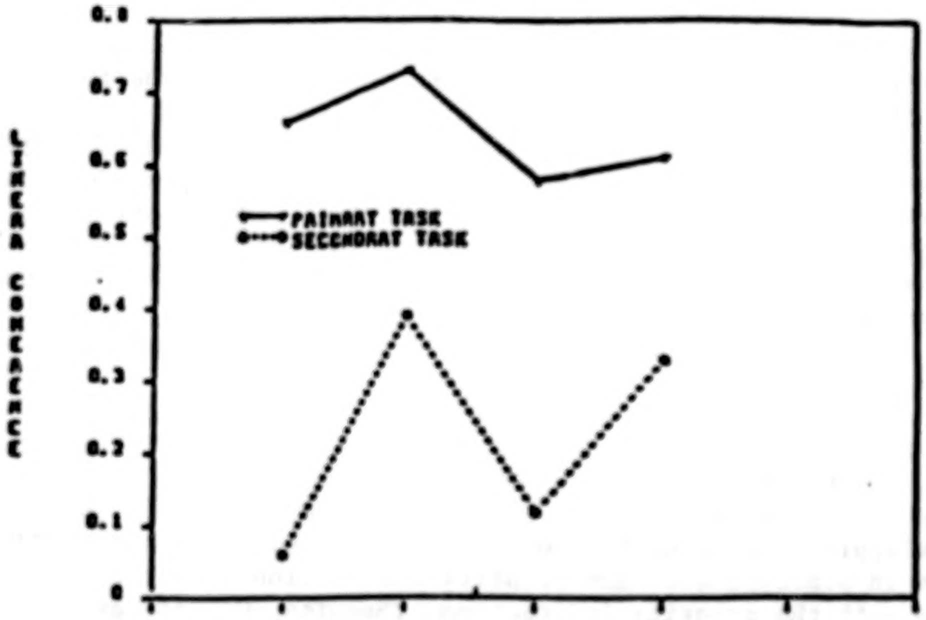


Figure 6. Linear coherence measures obtained between the tracking performances and the primary task difficulty in the dual tracking condition.

obtained from ensemble averages, no error terms were available to test statistically the interaction between the primary and secondary tasks.

The linear coherence measures obtained for the memory-tracking conditions were even more non-optimal. With the exception of the VM-TR condition (which showed the least change), the primary task coherence measures for the other three memory-tracking conditions generally increased between Sessions 7 (before the priority instructions were introduced) and 10 (after practice with the priority instructions).

Upon closer examination of the moment by moment tracking performance of each of the five dual task conditions, a further differentiation among the memory-tracking task pairs was observed. Recall that the subjects were instructed to maintain a constant primary task performance by allocating the appropriate amount of resources to the task as the difficulty level dictates. As such, any error spikes occurring at the increase of difficulty such as those found in Figure 5 can be considered as signs of resource allocation non-optimality. Despite the fact that the primary task was the same for all pairs of dual tasks, sharp primary task error spikes were found to be present only in the memory-tracking ensembles and not in the dual tracking ensembles obtained when the priority instructions were first introduced (Session 8). After some practice with the priority instructions (Session 10), the error spikes were much reduced for those memory-tracking pairs employing a manual secondary response (VM-TR and AM-TR) but not for those employing a speech secondary response (VS-TR and AS-TR).

To examine the error spikes data in a more quantitative fashion, the amplitude of the error spikes occurring at each of the four rising slopes of the primary task difficulty function was estimated by subtracting the value of the tracking error at the base of the spike from that at the peak of the spike. The mean spike amplitudes obtained from Sessions 8 and 10 are plotted in Figure 7. The structurally identical task pair (TR-TR) clearly has the smallest error spikes early in practice while the pair with the same I/O modalities but separate stages and codes of processing (VM-TR) has the next smallest error spikes. The pairs with separate stages and codes of processing and either separate input or separate output modalities (AM-TR or VS-TR) were found to have spikes of moderate magnitude. Finally, the error spikes are the largest for the task pair with completely separate stages and codes of processing and I/O modalities (AS-TR). Thus, it appears that the magnitude of the error spikes could be ordered by the degree of non-overlapping resources between the time-shared tasks and particularly so early in practice with the priority instructions. After some practice with the priority instructions (Session 10), the amplitudes of the error spikes for the VM-TR and AM-TR conditions were much reduced indicating the subjects' improved ability to guard their primary task performance against the momentary difficulty increases in these conditions. On the other hand, the magnitude of the error spikes of those memory-tracking pairs with separate output modalities remained large reflecting subjects' failure in protecting their primary task

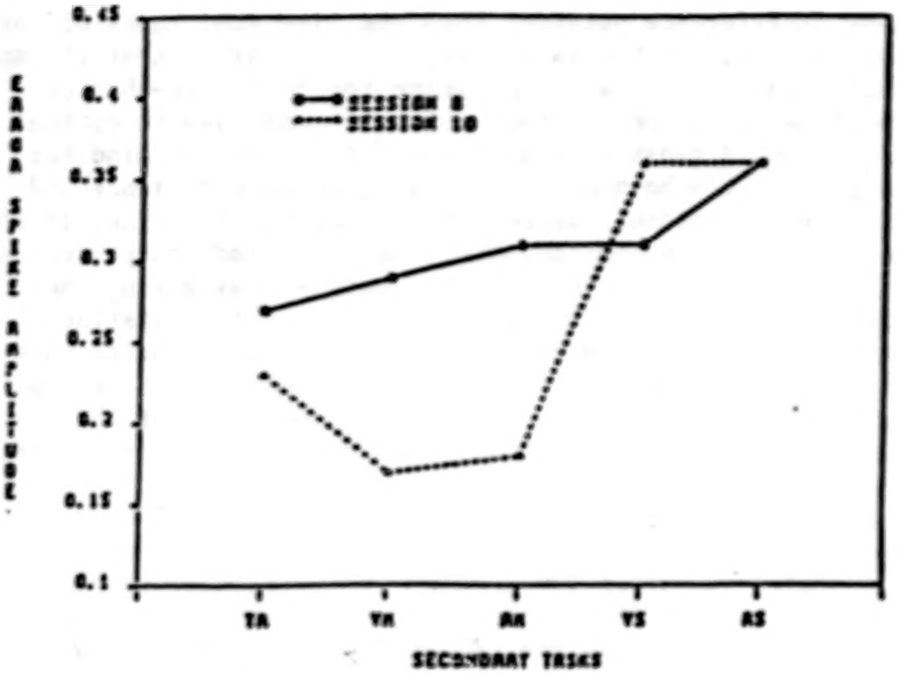


Figure 7. Error spikes amplitude (in RMSE) observed in the primary task ensemble averages of five structurally different task pairs.

performance from the difficulty increases even by the end of the experiment.

ANOVA results show that the session main effect was not significant but the task main effect was reliable at .01 level. Post hoc comparisons (Scheffe) show that the spike amplitudes (collapsed over sessions) for the AS condition were significantly larger than those for the VM condition with $p > .10$. The rest of the pairwise comparisons were not reliably different. Results from a separate ANOVA (Session x Input x Output) showed that the two input modalities were not significantly different from each other ($p > .10$) but the two output modalities were reliably different at .01 level. The Session x Output interaction was also significant ($p < .02$), confirming the observation that the error spikes obtained in the VM and AM ensembles in Session 8 were much reduced by Session 10, but the spikes in the VS and AS ensembles were unchanged by practice. In short, Figure 7 suggests that resource allocation optimality is related to the degree of overlapping resources between the time-shared tasks as hypothesized.

DISCUSSION

One of the task design guidelines suggested by the multiple resource theory is that time-shared tasks should be as structurally different as possible in order to maximize the total potential resources available and to minimize task interference. Such a recommendation is strongly supported by the present finding of systematic variation in the degree of task interference obtained from the five dual tasks of various structural similarity. At the same time, it is apparent that the same dual task configuration that would generate the best time-sharing efficiency would also be the one that is least conducive to optimal resource allocation. A question that comes immediately to mind is: Is there no design that can both maximize time-sharing efficiency and optimize resource allocation. There does not appear to be any if the different resources are indeed completely separate and independent of each other. In this case, one must choose between maximizing the performance of the time-shared tasks as a whole and facilitating resource allocation so that the performance of a high priority task will always be protected. The choice will naturally depends on the goal of the mission. However, the data presented here suggest that the relationship between the different resources may not be this simple.

First, the present results suggest that the degree of task interference between the time-shared tasks does not depend entirely on the number of shared resources but also on the particular common resources utilized by the time-shared tasks. In Figure 3, although the AM and VS memory tasks each has one I/O modality in common with the tracking task, the primary task error of the AM condition was found to be higher than that of the VS condition. This can perhaps be explained by the fact that the primary tracking task relies heavily on the response resources and thus competition between the tracking task and the manual memory task would be particularly severe (see also Vidulich & Wickens, 1981). What we need then is a reliable methodology and eventually a data base that could tell us a priori the precise structural composition of the task of interest.

Second, that the error spikes for the manual memory conditons could be reduced by practice but not those for the speech conditions (see Figure 7) posts the question of whether the three dimensions by which the processing resources are defined in the structure-specific resource model play an equal role in determining the degree of time-sharing efficiency and resource allocation optimality. For example, proper training may be able to reduce the resource allocation non-optimality between certain resources but not others. Further, as suggested by Wickens (1981), the different resources may not be all equally functionally distinct. Certain dimensions may have partially sharable resources (and hence amenable to training) while others are totally independent resources. Much research effort is still needed to unfold the functional organization of the multiple resources -- an understanding that will help reveal the precise circumstances under which the tradeoff between time-sharing efficiency and resource allocation optimality will occur and the extent to which such a tradeoff will occur.

ACKNOWLEDGEMENT

This research was supported by contract #N-000-14-79-C-0658 from the Office of Naval Research Engineering Psychology Program. Gerald Malecki was the technical monitor.

REFERENCES

Brickner, M. & Gopher, D. (1981, February). *Improving time-sharing performance by enhancing voluntary control on processing resources* (Technical Report AFOSR-77-3131C). Israel Institute of Technology, Research Center for Work Safety and Human Engineering.

Gopher, D. & Navon, D. (1980). How is performance limited: Testing the notion of central capacity. *Acta Psychologica*, *46*, 161-180.

Israel, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. (1980). P300 and tracking difficulty: Evidence for multiple resources in dual task performance. *Psychophysiology*, *17*, 57-70.

Kantowitz, B. H. & Knight, J. L. (1976). On experimenter-limited process. *Psychological Review*, *83*, 502-507.

Ogden, G., Levine, J., & Eisner, E. (1979). Measurement of workload by secondary tasks. *Human Factors*, *21*(5), 529-548.

Pierce, B. J. & Wickens, C. D. (1978). Linear Modelling of attentional resource allocation. *Proceedings of the 14th Annual Conference on Manual Control* (pp. 557-567).

Triesman, A. & Davies, A. (1973). Divided attention to eye and ear. In S. Kornblum (Ed.), *Attention and Performance IV*. New York: Academic Press.

Vidulich M. & Wickens, C. D. (1981, December). *Time-sharing manual control and memory search: The joint effects of input and output modality competition, priorities and control order* (Technical Report EPL-81-4/ONR-81-4). Champaign, Il: University of Illinois Engineering-Psychology Laboratory.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson & R. Pew (Eds.), *Attention and Performance VIII*. Englewood Cliffs, NJ: Lawrence Erlbaum.

Wickens, C. D. & Tsang, P. (1979). Attention allocation in dynamic environments. *Proceedings of the 15th Annual Conference on Manual Control* (pp. 82-92).

Wickens, C. D., Tsang, P., & Benel, R. (1979). The dynamics of resource allocation. In C. Bensel (Ed.), *Proceedings of the 23rd Annual Meeting of the Human Factors Society* (pp. 527-531). Santa Monica, CA: Human Factors.

Zeitlin, C. R. & Finkelman, J. M. (1975). Research note: Subsidiary task techniques of digit generation and digit recall as direct measures of operator loading. *Human Factors*, *17*(2), 218-220.

N85
14563

UNCLAS

D28

N85 14563

ON CHOOSING BETWEEN TWO PROBABILISTIC CHOICE SUB-MODELS IN A DYNAMIC MULTITASK ENVIRONMENT*

by

Eric P. Soulsby

Dept. of Electrical Engineering
and Computer Science
University of Connecticut
Storrs, CT 06268

ABSTRACT

Probabilistic decision theories are commonly divided into two types: constant utility models and random utility models. The two types of choice models differ in the locus of the random component. In the constant utility models the randomness is attributable to the decision rule, whereas in the random utility model the randomness is attributable to the utilities.

An independent random utility model based on Thurstone's Theory of Comparative Judgment and a constant utility model based on Luce's Choice Axiom are reviewed in detail. Predictions from the two models are shown to be equivalent under certain restrictions on the distribution of the underlying random process. Each model is applied as a stochastic choice submodel in a dynamic, multi-task, environment. Resulting choice probabilities are nearly identical, indicating that, despite their conceptual differences, neither model may be preferred over the other based solely on its predictive capability.

* Research supported in part under US Air Force Contract No. F33615-81-K-0510

PRECEDING PAGE BLANK NOT FILMED

I. INTRODUCTION AND MOTIVATION OF THE RESEARCH

Based on the Dynamic Decision Model (DDM) developed by Pattipati, et al [18-21] and the experimental paradigm used to validate it, the present paper explores the concepts behind the stochastic choice sub-model. Specifically, two different classes of probabilistic theories of choice, the *random utility* models and the *constant utility* models, are investigated in regard to a dynamic multi-task decisionmaking environment. In particular, a derivation of a random utility model based on *Thurstone's Law of comparative judgement* [22-24] will be presented, followed by a development of a constant utility model based on *Luce's Choice Axiom* [9-13]. This work was motivated by the desire to extend similarities exhibited between the two theories from that of paired comparisons in a static setting to that of multi-task, dynamic decision-making environments.

The organization of the paper is as follows: Part II describes the two classes of probabilistic theories of choice and elaborates on a representative example of each. Part III briefly reviews the Dynamic Decision Model and then discusses a comparison of the two approaches with regard to the dynamic multi-task environment. Model/data comparisons for each of the two approaches will be presented. Finally, Part IV comments on the results obtained.

II. PROBABILISTIC THEORIES OF CHOICE

It is well known that when faced with the same alternatives, under seemingly identical conditions, people do not always make the same choice [5]. Inconsistency is thus one of the basic characteristics of individual choice behavior. Fluctuations in response to the same stimulus occur even when there are no changes in the information or resources available to the decision maker. Therefore, one is led to the hypothesis that the observed inconsistency is a consequence of an underlying random process.

Probabilistic decision theories may be divided into two types: *constant utility* models and *random utility* models. Constant utility models assume that each alternative has a constant or fixed value and that the probability of choosing one alternative over another is a function of the distance between their utilities. Random utility models assume that the decision maker always chooses the alternative that has the highest utility, but the utilities themselves are random variables. Thus the actual choice mechanism is deterministic, but the utility of each alternative is subject to momentary fluctations.

The two types of choice representations are different in the locus of the random component. In the constant utility models the randomness is attributed to the decision rule, whereas in the random utility model the randomness is attributable to the utilities. Both models are closely related to psychophysical theories in which the probability of judging one object greater (heavier) than another is expressable as a monotonic function of the difference of their scale values. The decision problem is considered as a discrimination problem where the decisionmaker is trying to determine which alternative would be more satisfying [5].

Although these two types of choice representation are very different in psychological terms, they are somewhat compatible in mathematical terms. In the remainder of this section two probabilisitic theories of choice will be considered and will be shown to be quite similar mathematically.

Thurstone's Theory of Comparative Judgment

One of the first to postulate that choice behavior is probabilistic in nature was L. L. Thurstone [22-23] who noted that "... an observer is not consistent in his comparative judgments from one occasion to the next. He gives different comparative judgments on successive occasions about the same pair of stimuli". To account for the fluctations in the psychological evaluations of objects, Thurstone introduced the theory of comparative judgment.

The basic model underlying Thurstone's theory may be described by considering the two stimuli j and k shown in Fig. 1. When presented together to the observer, each stimulus excites a *discriminal process* d_j and d_k. The stimulus effects d_j, d_k are normally distributed with means s_j, s_k and *discriminal dispersions* σ_j, σ_k. The difference in discriminal processes $(d_k - d_j)$ for any single presentation of the pair of stimuli is called a *discriminal difference* and is denoted by d_{k-j}.

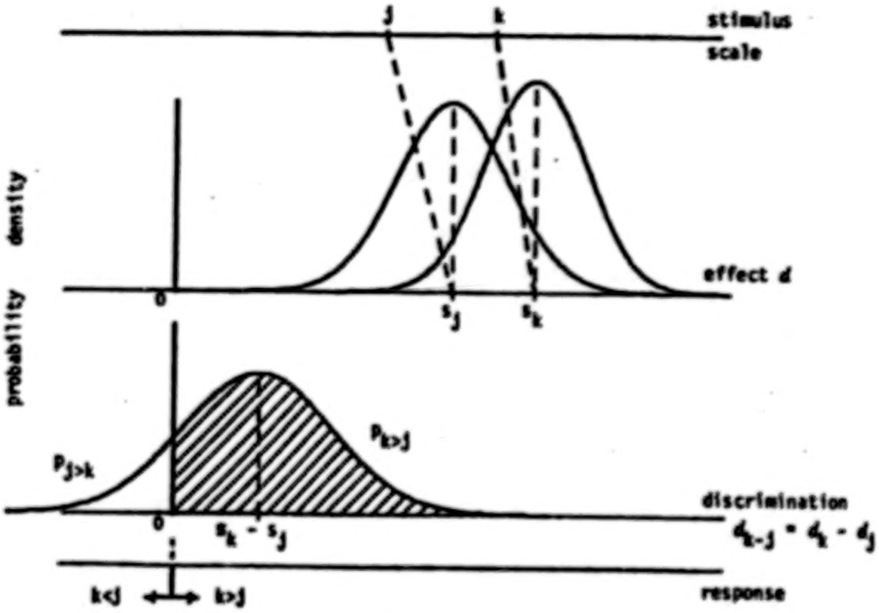


Fig. 1. VARIABLES RELEVANT TO THURSTONE'S THEORY

The probability density for d_{k-j} also has a normal distribution with mean $\bar{d}_{k-j}$, which is equal to the difference in scale values of the two stimuli, and standard deviation, given by

$$\sigma_{d_{k-j}} = (\sigma_j^2 + \sigma_k^2 - 2\rho\sigma_j\sigma_k)^{1/2}$$

where ρ represents the correlation between momentary values of discriminal processes associated with stimuli j and k.

As indicated in Fig. 1, it is assumed that the judgment "stimulus k is greater than stimulus j" occurs whenever the discriminal difference, d_{k-j}, is positive. Whenever this discriminal difference is negative, the judgment "stimulus j is greater than stimulus k" will be obtained. Thus the shaded portion to the right of the zero point gives the relative observed frequency of the judgment $k > j$. Under the assumption of normality the area under the curve for $d_{k-j} > 0$ on the discriminat axis equals

$$P_{k>j} = \frac{1}{\sigma_{d_{k-j}}\sqrt{2\pi}} \int_0^{\infty} \exp - \left\{ \frac{(x-\bar{d}_{k-j})^2}{2\sigma^2_{d_{k-j}}} \right\} dx$$

which by introducing

$$y = (x-\bar{d}_{k-j})/\sigma_{d_{k-j}}$$

becomes

$$P_{k>j} = \frac{1}{\sqrt{2\pi}} \int_{z_{kj}}^{\infty} \exp \{ - y^2/2 \} \, dy$$

in which $z_{kj} = - \bar{d}_{k-j}/\sigma_{d_{k-j}}$. Therefore, from the theoretical proportion of times stimulus k is judged greater than stimulus j, the value of z_{kj} may be determined from a table for the normal distribution. Since a relation between the standard deviation of the differences and the discriminal dispersions of the two stimuli is known, we arrive at the general form of the *law of comparative judgment*:

$$s_j - s_k = z_{kj} \sqrt{\sigma_j^2 + \sigma_k^2 - 2\rho\sigma_j\sigma_k}$$

Thurstone (1927b) presented several simplified forms of the law, of which case V involves the additional assumptions that the correlations between all pairs of stimuli are zero and that all the discriminal dispersions are equal, hence $s_j - s_k = z_{kj}\sigma$. If P_{jk} is defined as the probability that $j>k$, then case V may be expressed as

$$P_{jk} = \int_{-\infty}^{z_{kj}} \frac{1}{\sqrt{2\pi}} \exp - \left\{ \frac{y^2}{2} \right\} dy$$

In principle, it is rather simple to generalize Thurstone's model into a model for the choice of one stimulus from a set of more than two stimuli. This leads to the general form of the *random utility* models as follows [12]: Let A be the finite set of alternatives and let U be a function defined on A such that for each x in A, U(x) is a random variable. Then a random utility model is a set of preference probabilities defined for all subsets of a finite A for which there is a random vector U on A such that for $x \in Y \subseteq A$, the probability of choosing x from the set y is given by

$$P_Y(x) = \Pr\{U(x) \geq U(y)\ ,\ y \in Y\} = \int_{-\infty}^{\infty} \Pr\{U(x) = t\ ,\ U(y) \leq t\ ,\ y \in Y\}\ dt$$

If the definition is only asserted for the binary preference probabilities, that is, the probability of choosing x over y is $P(x,y) = \Pr\{U(x) \geq U(y)\}$ then the model is called a *binary* random utility model. If the random vector U consists of components that are independent random variables, then we say the model is an *independent* random utility model, in which

$$P_Y(x) = \int_{-\infty}^{\infty} \Pr\{U(x) = t\} \prod_{y \in Y - \{x\}} \Pr\{U(y) \leq t\}\ dt$$

Therefore, generalizing Thurstone's theory into an independent random utility model leads to the following: Let $P_A(x)$ be the probability that alternative x is chosen from the set $A = \{x,y,\ldots\}$. By assuming, along the lines of Thurstone, that the sensation of stimulus x has a normally distributed discriminal process with mean $\bar{x}$ and variance σ_x^2, then

$$\Pr\{U(x) = t\} = \frac{1}{\sqrt{2\pi}\ \sigma_x} \exp - \left\{\frac{(t-\bar{x})^2}{2\sigma_x^2}\right\}$$

and

$$\Pr\{U(y) \leq t\} = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\ \sigma_y} \exp - \left\{\frac{(u-\bar{y})^2}{2\sigma_y^2}\right\} du$$

which, upon a change of variables along with letting Φ represent the standard normal integral, leads to

$$P_A(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}}\ e^{-z^2} \prod_{\substack{y \in A \\ y \neq x}} \left[\Phi\left(\frac{\sqrt{2}\sigma_x z + \bar{x} - \bar{y}}{\sigma_y}\right)\right] dz$$

which can be evaluated numerically via Gauss-Hermite quadrature [27].

Luce's Choice Axiom [9]

A second probabilistic model, applicable to paired comparisons as well as to choices from more than two alternatives, is a random response model based on Luce's Choice Axiom [9]. Luce accepted the basic assumption that choice behavior is governed by a random process, but instead of making additional assumptions about the form of the value distribution he assumed that choice probabilities satisfy a simple, but powerful, axiom that serves as a cornerstone of the model.

The choice axiom, in essence, states that the removal of some alternatives from the choice set A does not alter the relative probabilities of choice among the remaining alternatives. In other words, the presence or absence of an alternative is irrelevant to the relative probabilities of choice between two other alternatives, although the absolute values of these probabilities will in general be affected. Formally, for all $x \in A \subseteq T$ Luce's Choice Axiom asserts that the probability of choosing an element x of A, from the entire set T, $P_T(x)$, equals the probability that the selected alternative will be in the subset A, $P_T(A)$, multiplie by the probability of choosing x from A, $P_A(x)$. Mathematically,

$$P_T(x) = P_A(x)P_T(A) \qquad \text{for } A \subseteq T$$

The axiom states that the probability of selecting, for example, roast beef (x) from a menu (T) equals the probability of selecting roast beef from the meat entree (A) times the probability of choosing a meat entree. Stated differently, the axiom implies that the probability of choosing x from A equals the conditional probabilit of choosing x from T given that the choice is restricted to A, i.e. $P_A(x) = P_T(x|A)$

Using the choice axiom and the fundamental laws of probability theory, Luce derived a large number of interesting consequences, some of which are presented in the following theorem (for proof see [9, 28]):

Theorem: If for all $x \in T$, $P_T(x) \neq 0$ and if the choice axiom holds for all x and A such that $x \in A \subseteq T$, then, by letting $p(x:y) = P_{\{x,y\}}(x)$ = prob. of choosing x when presented with the two alternatives x and y, we obtain

$$\text{(i)} \qquad \frac{p(x:y)}{p(y:x)} = \frac{P_T(x)}{P_T(y)} = \frac{P_A(x)}{P_A(y)}$$

and

$$\text{(ii)} \qquad p_T(x) = \left[1 + \sum_{y \in T - \{x\}} \frac{p(y:x)}{p(x:y)}\right]^{-1}$$

The result can be extended to any subset $A \subseteq T$ that contains the alternatives x and y. The condition in (i) is commonly called the constant ratio rule which expresses the fact that, under the choice axiom, the ratio of the form $P_R(x)/P_R(y)$ is independent of R. That is, the ratio of the probability of selecting steak for dinner and the probability of selecting roast beef for dinner is the same for all menus containing both entrees. This rule may also be considered as a probabilistic version of the principle of independence from irrelevant alternatives, i.e. "our preferences between specific objects do not change when other objects are added to, or subtracted from, the overall set of objects" [17].

Comparing the theories of Thurstone and Luce

Based on the choice axiom, Luce [9] derived a relationship among triples of pairwise probabilities which he used as a basis for comparing his model with Case V of the Thurstone model. Luce [9] presented results showing a comparison of predicted p(x:z) from known p(x:y) and p(y:z) shown in the following table:

TABLE 1. Comparison of Predicted p(x:z) from Known p(x:y) and p(y:z) Using Axiom 1 and Thurstone's Case V. (from Luce [9])

		p(y:z)			
		0.6	0.7	0.8	0.9
p(x:y)	0.6	0.692	0.778	0.857	0.931
	0.7		0.845	0.903	0.954
	0.8			0.941	0.973
	0.9				0.988

p(x:z) from axiom 1

		p(y:z)			
		0.6	0.7	0.8	0.9
p(x:y)	0.6	0.695	0.782	0.864	0.938
	0.7		0.853	0.915	0.965
	0.8			0.954	0.983
	0.9				0.995

p(x:z) from Thurstone's case V

Thus, "although the two models are based on different assumptions, they predict practically the same values and hence it is quite difficult to choose between them on empirical grounds" [9].

Holman and Marley (cited in Luce and Suppes [12]) showed that if Thurstone's discriminal processes are assumed to have the *double exponential distribution*

$$F(x) = e^{-e^{-(ax+b)}} \qquad (a > 0,\ b \text{ arbitrary})$$

then the Thurstone model is completely equivalent to the Choice Axiom. More recently, in a comprehensive comparison between the two theories, Yellot [26] showed that for paired comparisons the relation, cited above, between Thurstone's Case V and Luce's Choice Axiom is not unique; i.e. other discriminal process distributions also yield a model equivalent to the Choice Axiom. Yellot, however, proved in general the following theorem (for proof see [26]):

Theorem: A Thurstone model [case V] is equivalent to the Choice Axiom for complete experiments [one in which choice probabilities are determined for every subset of objects] with three (or more) objects if, and only if, [the associated Thurstone model's distribution] F is a double exponential distribution.

Fig. 2 shows a comparison of discriminal process probability density functions for two different models:

(a) the Normal (Case V) model:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp - \left\{ \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\} \quad , \quad \mu=0 \quad , \quad \sigma=1$$

and (b) the Double Exponential model:

$$f(x) = \frac{1}{\theta} e^{-\frac{(x-\xi)}{\theta}} \exp\left\{ -e^{-\frac{(x-\xi)}{\theta}} \right\} \quad , \quad \xi=0 \quad , \quad \theta=\sqrt{6}/\pi$$

The mean, mode, and median of the Normal density is given by μ, and its variance by σ^2. The Double Exponential distribution has its mode at ξ, its median at

$\xi-\theta\log\log 2$, its mean at $\xi+\gamma\theta$ (where γ = Euler's constant = .57722), and its variance is given by $\frac{1}{6}\pi^2\theta^2$ [7].

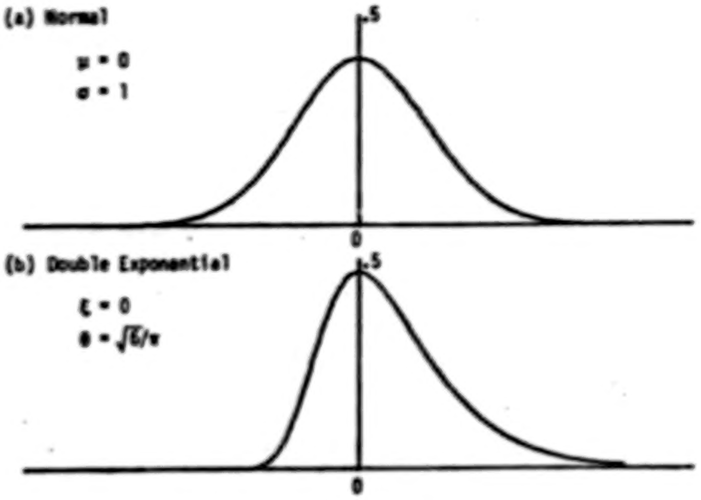


Fig. 2. DISCRIMINAL PROCESS PROBABILITY DENSITIES

III. APPLICATION IN A DYNAMIC MULTI-TASK ENVIRONMENT

Review of the Dynamic Decision Model

A block diagram of the Dynamic Decision Model (DDM) developed by Pattipati, et al [18-21] for modeling human decisionmaking performance in a dynamic multi-task environment is shown in Fig. 3. The basic assumption underlying the development of the dynamic decision model is that a well-trained human behaves in a normative, rational manner subject to his inherent limitations. Mathematically this may be interpreted in terms of maximizing a specified metric. Pattipati, et al [18-2 utilized the subjectively expected value (SEV) of a decision as a metric (or "attractiveness measure") for optimization.

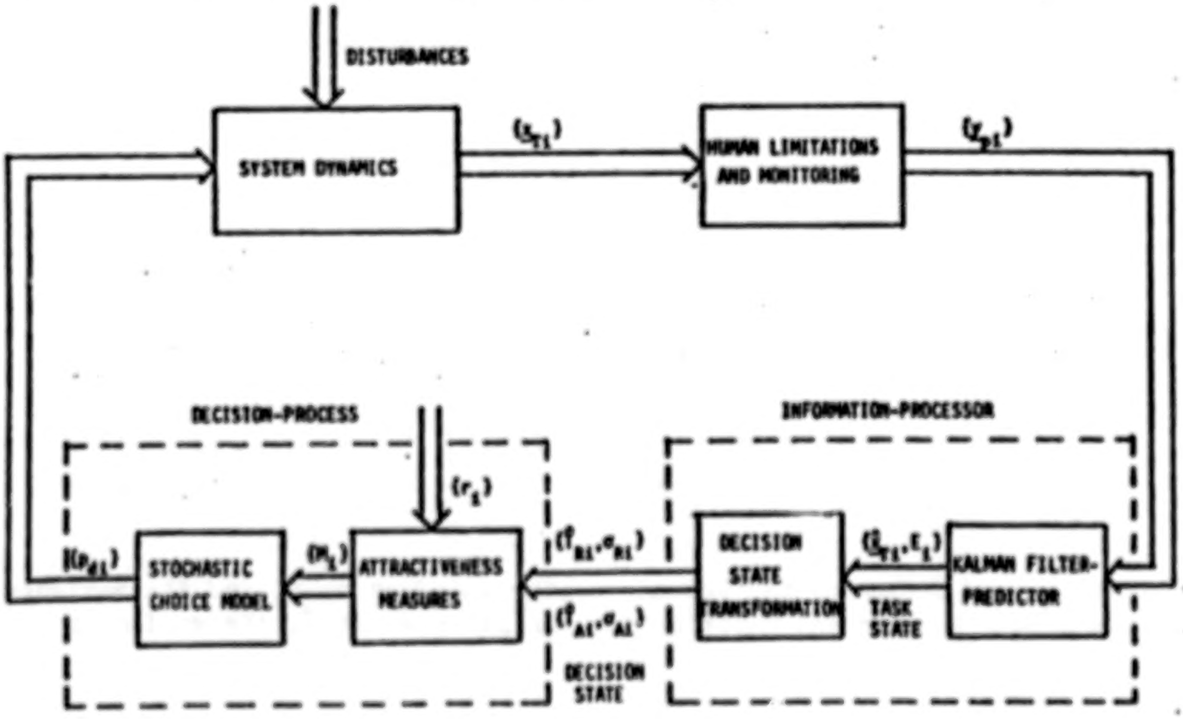

DISTURBANCES
SYSTEM DYNAMICS
(x_{Ti})
HUMAN LIMITATIONS AND MONITORING
(y_{pi})
DECISION-PROCESS
INFORMATION-PROCESSOR
(r_i)
(p_{di})
STOCHASTIC CHOICE MODEL
(M_i)
ATTRACTIVENESS MEASURES
$(\hat{f}_{Ri},\sigma_{Ri})$
$(\hat{f}_{Ai},\sigma_{Ai})$
DECISION STATE
DECISION STATE TRANSFORMATION
$(\hat{x}_{Ti},\Sigma_i)$
TASK STATE
KALMAN FILTER-PREDICTOR

Fig. 3. DYNAMIC DECISION MODEL OF HUMAN TASK SELECTION PERFORMANCE

The decision strategy incorporated into the DDM may be summarized as follows (see [18-21,28] for details): Each of the N tasks in the opportunity window is represented by a dynamic system acted on by disturbances to account for the non-stationarities in task characteristics. The human's perceived outputs $\{\underline{y}_{pi}\}$ are delayed, noisy versions of the task states $\{\underline{x}_{Ti}\}$ and are contingent upon the monitoring process. The perceived outputs are processed to produce the best linear unbiased estimates of the task states $\{\underline{x}_{Ti}\}$ and their associated covariances $\{E_i\}$ via a Kalman filter-predictor submodel. The statistics of the task states $\{\hat{x}_{Ti}, E_i\}$ are, in turn, used to determine the first and second order statistics of the decision state $\{T_{Ri}, \sigma_{Ri}\}$ time required and $\{T_{Ai}, \sigma_{Ai}\}$ time available. The combined statistics of the decision states of the tasks in the opportunity window are used to compute the transition probabilities among the various process states for each of the decision alternatives. The transition probabilities, along with the task values, are used to determine the attractiveness measures of the tasks, employing the SEV criterion. These measures form the input to a stochastic choice model which generates the decision probabilities. Fig. 4 depicts the human decision process as modelled by the DDM.

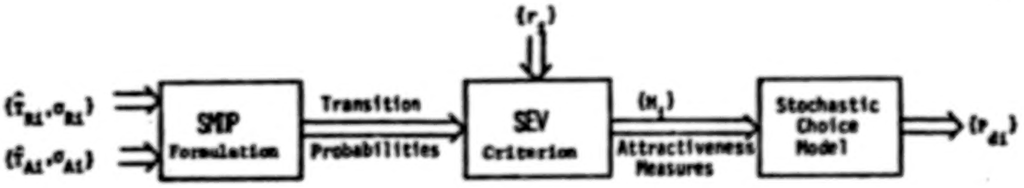

$\{r_i\}$
$\{\hat{T}_{Ri}, \sigma_{Ri}\}$
$\{\hat{T}_{Ai}, \sigma_{Ai}\}$
SMP
Formulation
Transition
Probabilities
SEV
Criterion
$\{M_i\}$
Attractiveness
Measures
Stochastic
Choice
Model
$\{P_{di}\}$

Fig 4. HUMAN DECISION PROCESS

The possible transition events associated with a decision to act on a task i in a process state $\underline{s}$ at time t are shown in Fig. 5. The event, "successful completion of task i", occurs if the decision state variable $T_{Ri}(t)$ of task i is greater than zero but less than the available times, $T_{Ai}(t)$, of all the tasks including i in the opportunity window. The probability of this event, denoted by $\eta_i(t)$ is given by

$$\eta_i(t) \triangleq P \text{ \{action on task i, task i is successfully completed, other tasks intact\}}$$

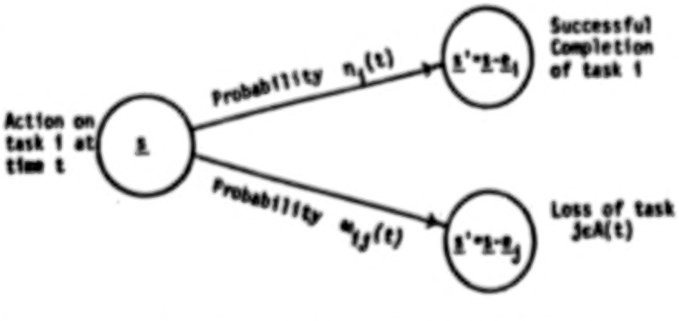

Action on
task i at
time t
$\underline{s}$
Probability $\eta_i(t)$
$\underline{s}' = \underline{s} - \underline{e}_i$
Successful
Completion
of task i
Probability $\omega_{ij}(t)$
$\underline{s}' = \underline{s} - \underline{e}_j$
Loss of task
$j \in A(t)$

Fig. 5. TRANSITION EVENTS FOR THE MULTI-TASK DECISION PROBLEM

On the other hand, a task j is said to be "lost" (while acting on task i) if $T_{Aj}(t)$ is greater than zero, but is less than $T_{Ri}(t)$ and $T_{Am}(t)$, $m \neq j$. The probability of this event, $\omega_{ij}(t)$, is

$$\omega_{ij}(t) \triangleq P\{\text{action on task } i\text{, an accessible task } j \text{ is lost, all the other tasks intact}\}$$

The attractiveness measure, $M_i(t)$, of a decision to act on task i is given by

$$M_i(t) = r_i(t)\eta_i(t) - \sum_{j\in A(t)} q_j(t)\omega_{ij}(t)$$

where $A(t)$ represents the set of available tasks in the opportunity window at time t, $r_i(t)$ are the task values (rewards) and $q_j(t)$ are the (subjective) losses assigned by the decisionmaker for the loss of task j. In the current DDM, the subjective $q_j(t)$ is the objective value, $r_j(t)$.

Stochastic Choice Sub-models

The stochastic choice model used by Pattipati, et al [18-21] was based on Luce's Choice Axiom which has been discussed at length in Part II. In order to make the axiom fit into the decision framework, it was necessary to assume that although the attractiveness measures, $M_i(t)$, could be characterized by a single fixed number, the subjects perceive it as a random variable, $\tilde{M}_i(t)$, with an assumed Gaussian distribution. Pattipati, et al [20] indicate that "the randomness may be interpreted in terms of the uncertainties associated with the human perception of task values, $r_i(t)$, or his estimates of the transition probabilities, $\eta_i(t)$ and $\omega_{ij}(t)$".

Using Luce's Choice Axiom, the decision probabilities, $P_{di}(t)$, may be computed via

$$P_{di}(t) = \left[1 + \sum_{\substack{k\in\mathcal{D}(t)\\ k\neq i}} \frac{P\{\tilde{M}_k(t)-\tilde{M}_i(t) > 0\}}{P\{\tilde{M}_i(t)-\tilde{M}_k(t) > 0\}}\right]^{-1} \; ; \quad i\in\mathcal{D}(t)$$

where $\mathcal{D}(t)$ is the set of feasible decisions at time t. It is assumed that the $\tilde{M}_i(t)$ are Gaussian random variables with mean $M_i(t)$ and variance $\sigma^2_{\tilde{M}i}(t)$ that scales with $M_i^2(t)$. That is,

$$\sigma_{Mi}(t) = c|M_i(t)| \quad (c \approx .2\text{-}.4)$$

where c is the co-efficient of variation, a model parameter. The form of the above equation was motivated by the Weber's law of scaling in psychophysics.

Similarly, using the same approach as Pattipati, et al, one can envision the use of an independent random utility choice sub-model based on Thurstone's theory of comparative judgment. Specifically, we may make the same assumptions about the attractiveness measures as did Pattipati, et al. In fact, this assumption is none other than that originally proposed by Thurstone [22-24], i.e. the attractiveness measure M_i gives rise to a discriminal process which is normally distributed with mean $M_i(t)$ and standard deviation $c|M_i(t)|$.

Therefore, following the development in Part II, the generalized independent random utility model, based on Thurstone's theory of comparative judgment, can be applied to yield the decision probabilities, $P_{di}(t)$, as follows:

$$P_{di}(t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-z^2} \prod_{\substack{k \in \mathcal{D}(t) \\ k \neq i}} \Phi\left(\frac{\sqrt{2}\, \sigma_{M_i}(t) z + M_i(t) - M_k(t)}{\sigma_{M_k}(t)} \right) dz$$

Model Predictions

The dynamic decision model can be used in a straightforward manner to generate predictions of $P_{di}(t)$, as well as of other response measures that can be computed from the experimental data [18-21]:

The Completion Probability:

$P_{ci}(t)$ is the probability that task i is completed by time t.

The Error Probability:

$P_e(t)$ is the probability that the human commits an error, i.e., starts acting on a task he can not possibly complete.

The Average Accumulated Reward:

$\bar{R}(t)$ is the average total reward earned up to the present time t. It is an overall response measure.

Normalized Incremental Reward:

$W_c(t)$ is the average instantaneous reward-earning rate, and is a measure of instantaneous performance.

Total expected tasks completed:

$\bar{N}_c$ is the average number of tasks completed.

Average Time Spent On A Task On Line i:

$\bar{T}_{ij}(t)$ is the average time spent on a task on line i during the j-th pass.

DDM Results

Table 2 provides a comparison of the model predictions using the two stochastic choice sub-models with that of the experimental data. The ensemble data were obtained by averaging over 32 runs of a multi-task experiment (see [28] for details). As is readily seen, both stochastic choice submodels perform well in terms of these overall performance measures.

Table 2.

DDM RESULTS

	CONSTANT UTILITY (LUCE)	RANDOM UTILITY (THURSTONE)	DATA
EXPECTED REWARD EARNED:	56.596	56.719	58.125
TOTAL % REWARD EARNED:	75.461	75.626	77.500
EXPECTED TASKS COMPLETED:	24.801	24.862	28.000
TOTAL % TASKS COMPLETED:	63.591	63.75	71.795

The time histories of the action probabilities, as well as of the error probability, and the normalized incremental reward are given in Figs. 6-8 for each of the two stochastic choice sub-models. The model-data match is uniformly good to excellent for both models under study here. In fact, the action probabilities, $P_{di}(t)$, for both the constant utility model (Luce) and the random utility model (Thurstone) show remarkable similarities. This is most noteworthy, although perhaps not too unexpected.

Measures of Performance Similarity: Model/Data

In order to assess the closeness of model vs. data results, Pattipati, et al [18-21] developed several time-history and scalar measures of similarity. The time-history metrics compare the ensemble-averaged time history of a response variable obtained empirically with that predicted by the DDM. In the present multi-task paradigm, comparison of decision probabilities, completion probabilities, normalized incremental reward, accumulated reward and the error probability, appear to be five suitable time-history metrics.

Assessment of model-data similarity on the basis of time-history measures is often subjective, and is difficult to quantify. This resulted in the development of the following five scalar measures in the model-data validation studies (for details see [18-21,28]):

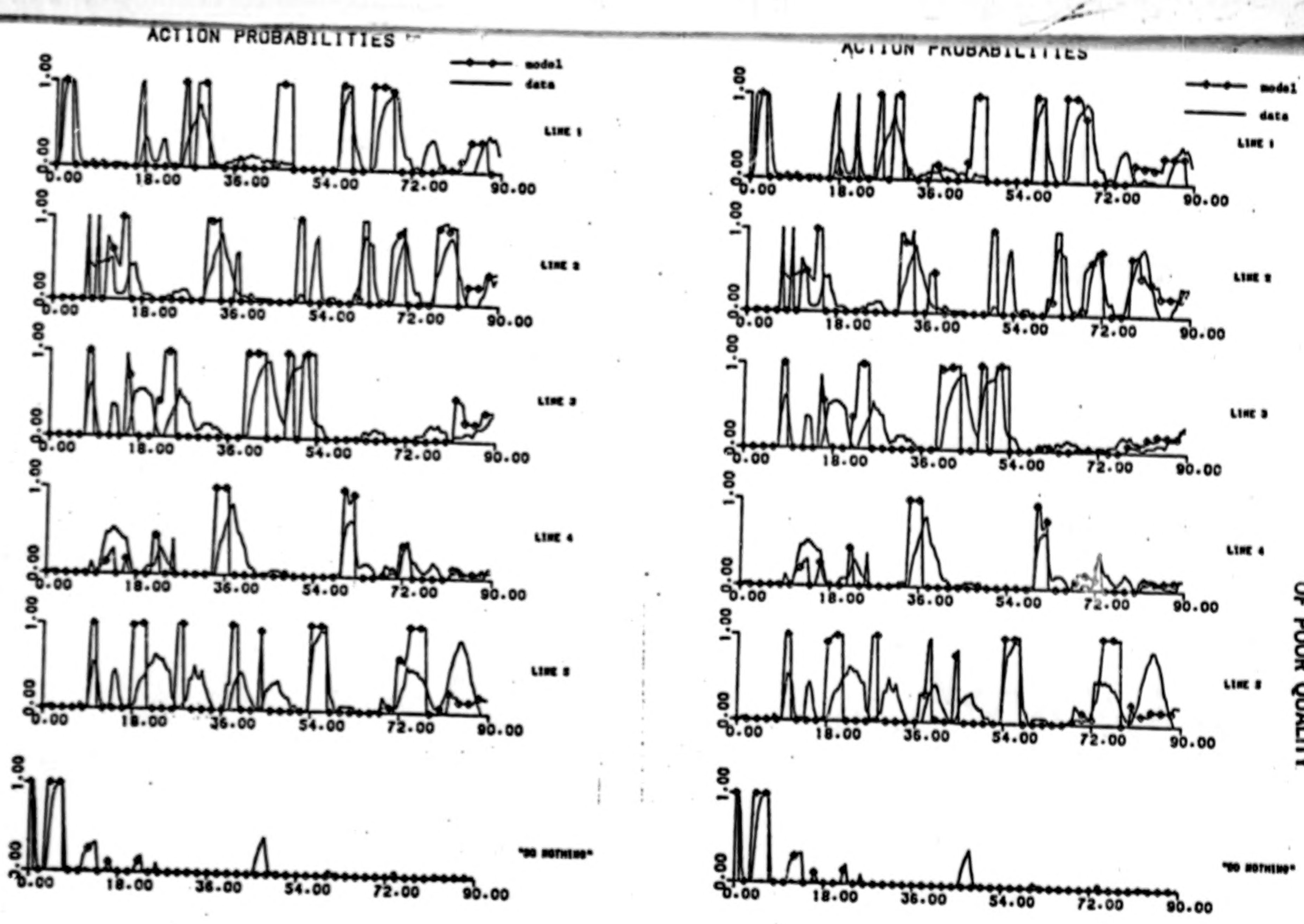
Action Metric:

AM computes the normalized time integral of the squared error differences between the decision probabilities.

Incremental Reward Metric:

IRM is the normalized time integral of the squared, weighted difference of

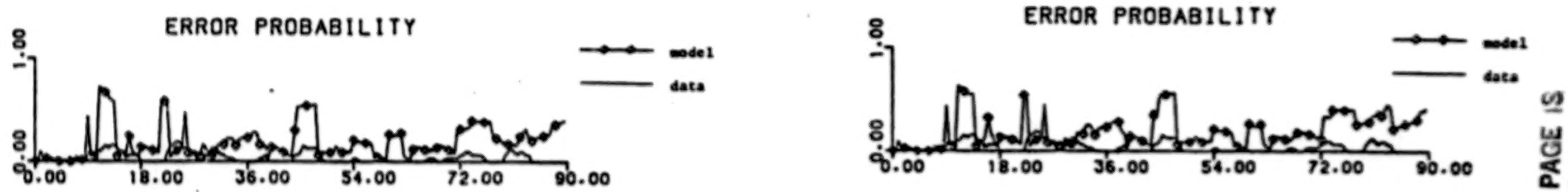
ORIGINAL PAGE IS
OF POOR QUALITY



(a) Luce choice sub-model

Fig. 6

(b) Thurstone choice sub-model

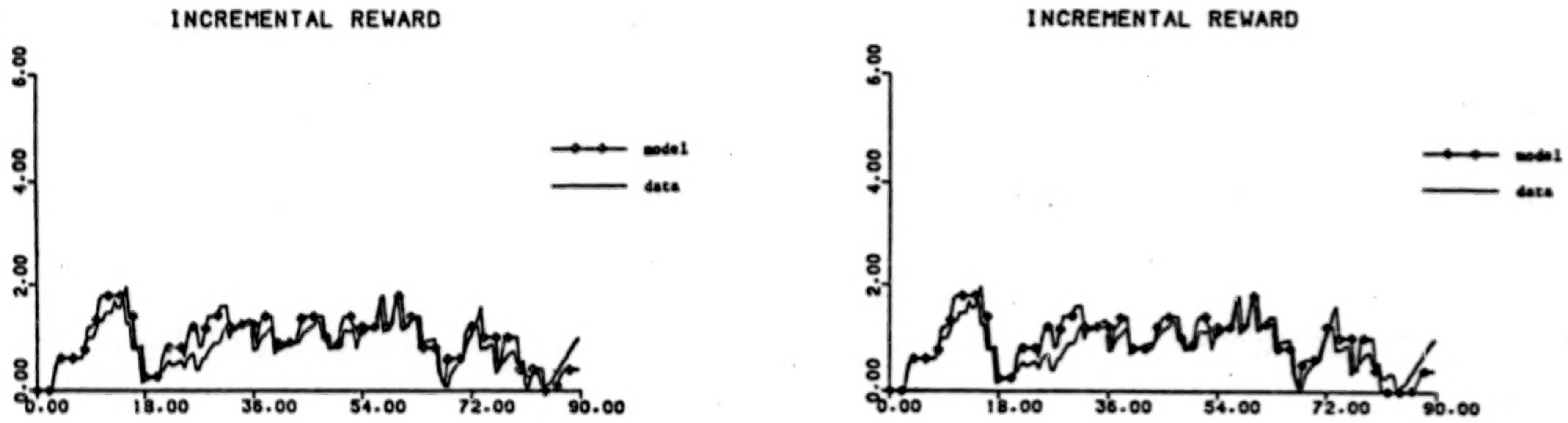


(a) Luce choice sub-model

Fig. 7

(b) Thurstone choice sub-model

ORIGINAL PAGE IS
OF POOR QUALITY



(a) Luce choice sub-model

Fig. 8

(b) Thurstone choice sub-model

the completion probabilities of the model and the data.

Accumulated Reward Metric:

ARM is the normalized time integral of the squared difference between the average reward earned up to that instant of time by the human and the model.

Average Time On Each Task Metric:

ATTM calculates the normalized root-mean-squared sum of the difference between the times spent on each task by the human and the model.

Error Probability Metric:

EPM is the normalized time integral of the squared differences between the error probabilities of the model and the data.

Note that the normalized scalar measures can range from a value of 0, corresponding to a perfect fit between the model and data, to a maximum value of 1. Table 3 provides a summary of the scalar measures of similarity for each of the stochastic choice submodels:

Table 3. MODEL/DATA PERFORMANCE METRICS

		CONSTANT UTILITY (LUCE)	RANDOM UTILITY (THURSTONE)
(1)	ACTION METRIC =	.095634	.095624
(2)	INCREMENTAL REWARD METRIC =	.10698	.10706
(3)	ACCUMULATED REWARD METRIC =	.0001791	.0001778
(4)	AVER. TIME ON EACH TASK METRIC =	.075815	.076635
(5)	ERROR PROBABILITY METRIC =	.055437	.055316

Stochastic Choice Sub-Model Comparison

As indicated in the preceding time history plots, and summarized in Table 3, both stochastic choice sub-models achieve similar matches to the experimental data. Indeed the decision probabilities predicted by each sub-model are nearly identical. This was not totally unexpected, since the predictions obtained from each sub-model have been shown to yield similar results under certain restrictions on the underlying distribution of the discriminal processes. Specifically, it has been shown [26] that if the underlying distribution is of the double exponential type, then both models are identical. Due to the similarities between the double exponential and the normal distribution (recall Fig. 2), it is not too surprising to find such a good agreement in the dynamic multi-task paradigm between the two sub-models as evidenced by the DDM predictions.

While we realize that the data used for the model/data comparison represents a "nominal" condition, from which broad conclusions may be mistakenly drawn, we feel that the agreement between the two approaches (random utility or Thurstone and constant utility or Luce) is indeed noteworthy. In fact, this agreement has been noted under various other experimental contexts.

IV CONCLUSIONS

The present paper has examined two dominant probabilistic theories concerning individual choice behavior. In particular, an independent random utility model based on Thurstone's theory of comparative judgment [22-24] and a constant utility model based on Luce's Choice Axiom [9] have been presented in detail. Similarities and differences between the two approaches have been discussed with emphasis placed on both the psychological and the mathematical aspects of each. Each theory has been applied as a stochastic choice sub-model in the Dynamic Decision Model (DDM) which was recently developed for modeling the decision maker's performance in a dynamic multi-task environment. Results have indicated that neither approach should be preferred over the other based strictly on the observed model/data match. One may note that although the two approaches are based upon two distinct philosophi regarding individual choice behavior, the results obtained in the present context are for the most part the same.

Future Directions

As mentioned previously, the choice probabilities obtained from either choice model are directly influenced by the set of alternatives presented to the decision maker. According to Luce's Choice Axiom, the presence or absence of an alternative is irrelevant to the relative probabilities of choice between two other alternatives although the absolute values of these probabilities will in general be affected. One aspect of operative behavior under stressful situations is that of reducing the number of alternatives that may reside in the decision maker's action set by some form of filtering mechanism. It appears (from preliminary work along these lines) that a mechanism by which elements may be discarded or added to the decision maker's action set would comprise a necessary addition to the two existing sub-models to account for this operative behavior under stressful situations. This remains a future direction of research.

REFERENCES

[1] Becker, G. M., DeGroot, M. H. and Marschak, J.: "Stochastic Models of Choice Behavior", Behavioral Science, vol. 8, 1963, pp. 41-55.

[2] : "An Experimental Study of Some Stochastic Models for Wagers", Behavioral Science, vol. 8, 1963, pp. 199-202.

[3] : "Probabilities of Choices Among Very Similar Objects: An Experiment to Decide Between Two Models", Behavioral Science, vol. 8, 1963, pp. 306-311.

[4] Burke, C. J. and Zinnes, J. L.: "A Paired Comparison of Pair Comparisons", J. of Mathematical Psychology, 2, 1965, pp. 53-76.

[5] Coombs, C. H.; Dawes, R. M. and Tversky, A.: Mathematical Psychology, An Elementary Introduction, 1970 by Prentice-Hall, Inc., Englewood Cliff, New Jersey.

[6] Debreu, G.: "Review of R. D. Luce's Individual Choice Behavior, A Theoretical Analysis", The American Economic Review, vol. L, Number 1, March 1960, pp. 186-188.

[7] Johnson, N. L. and Kotz, S.: Distributions in Statistics: Continuous Univariate Distributions - 1, 1970 by N. L. Johnson and S. Kotz, Houghton Mifflin Company.

[8] Kleinman, D. L.; Soulsby, E. P. and Pattipati, K. R.: "Decision Aiding - An Analytic and Experimental Study in a Multi-task Selection Paradigm", Proc. of the Fourth MIT/ONR Workshop on Distributed Information and Decision Systems Motivated by Command-Control-Communication (C^3) Problems, MIT, LIDS-R-1159, Oct. 1981.

[9] Luce, R. D.: Individual Choice Behavior, A Theoretical Analysis, 1959 by John Wiley & Sons, Inc., NY.

[10] Luce, R. D.: "Response Latencies and Probabilities", in Mathematical Methods in the Social Sciences, 1959 (Proc. of the First Stanford Symposium on Mathematical Methods in the Social Sciences, Stanford University, 1959) Edited by K. J. Arrow, S. Karlin, and P. Suppes, 1960 by Stanford Univerity Press.

[11] Luce, R. D. and Galanter, E.:"Discrimination", in Handbook of Mathematical Psychology, Edited by R. D. Luce, R. R. Bush, and E. Galanter, 1963 by John Wiley & Sons, Inc., vol. I, pp. 191-243.

[12] Luce, R. D. and Suppes, P.: "Preference, Utility, and Subjective Probability", in Handbook of Mathematical Psychology, Edited by R. D. Luce, R. R. Bush, and E. Galanter, 1965 by John Wiley & Sons, Inc., vol. III, pp. 249-410.

[13] Luce, R. D.: "The Choice Axiom After Twenty Years", Journal of Mathematical Psychology, 15, 1977, pp. 215-233.

[14] Marley, A. A. J.: "The Relation Between the Discard and Regularity Conditions for Choice Probabilities", Journal of Mathematical Psychology, 2, 1965, pp. 242-253.

[15] Marley, A. A. J.: "Some Probabilistic Models of Simple Choice and Ranking", Journal of Mathematical Psychology, 5, 1968, pp. 311-332.

[16] Michon, J. A.; Eijkman, E. G. J. and deKlerk, L. F. W.: Handbook of Psychonomics, 1979 by North Holland Publishing Company, NY.

[17] Morgan, B. J.T.: "On Luce's Choice Axiom", Journal of Mathematical Psychology, 11, 1974, pp. 107-123.

[18] Pattipati, K. R.; Ephrath, A. R. and Kleinman, D. L.: "Analysis of Human Decision-Making in Multi-task Environments", University of Connecticut Technical Report EECS-TR-79-15, Nov. 1979.

[19] Pattipati, K. R.: "Dynamic Decision-Making in Multi-Task Environments: Theory and Experimental Results", Ph.D. dissertation, Aug. 1980, University of Connecticut Technical Report EECS-TR-81-9.

[20] Pattipati, K.R.; Kleinman, D. L. and Ephrath, A. R.: "A Dynamic Decision Model of Human Task Selection Performance", University of Connecticut Technical Report EECS-TR-82-1, March 1982.

[21] Pattipati, K. R.; Kleinman, D. L. and Ephrath, A. R.: "A Dynamic Decision Model of Human Task Selection Performance", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, No. 2, March/April 1983, pp. 145-166.

[22] Thurstone, L. L.: "Psychophysical Analysis", *American Journal of Psychology*, 38, 1927a, pp. 368-389.

[23] : "A Law of Comparative Judgment", *Psychological Review*, 34, 1927b, pp. 273-286.

[24] : "*The Measurement of Values*", 1959 by the University of Chicago Press.

[25] Torgerson, W. S.: *Theory and Methods of Scaling*, 1958 by John Wiley & Sons, I

[26] Yellot, J. I.: "The Relationship Between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution", *Journal of Mathematical Psychology*, 15, 1977, pp. 109-144.

[27] *CRC Handbook of Tables for Mathematics*, Cleveland, Ohio. The Chemical Rubber Company, 1970.

[28] Soulsby, E. P.: "On Choosing Between Two Probabilistic Choice Sub-Models in a Dynamic Multitask Environment". University of Connecticut Technical Report EECS-83-11, July 1983.

Measurement of Workload and Performance in Simulation

N85
14564

UNCLAS

N85 14564 D29

NO FATIGUE EFFECT ON BLINK RATE

Wonsoo Kim, Wolfgang Zangemeister* and Lawrence Stark

Departments of Electrical Engineering and Computer Science, and Physiological Optics, University of California, Berkeley.

* Department of Neurology, University of Hamburg, Germany.

ABSTRACT

Blink rate has been reported to vary dependent upon ongoing task performance, perceptual, attentional and cognitive factors, and fatigue.

For our experiment, we operationally defined five levels of task difficulty and measured task performance as lines read aloud per minute. A single non-invasive infrared TV eyetracker was modified to measure blinking and an on-line computer program identified and counted blinks while the subject performed the tasks. Blink rate decreased by 50 % when either task performance increased (fast reading) or visual difficulty increased (blurred text); blink rate increased greatly during rest breaks.

There was no change in blink rate during one hour experiments even though subjects complained of severe fatigue.

INTRODUCTION

Fatigue. Fatigue is one of major factors of pilot workload, relevant both to safe control of airplane flight and also for long term well-being of pilots. However, fatigue is a psychophysiological phenomenon, difficult to measure or quantify. Three operational measures of fatigue can be considered --- subjective fatigue, general physiological factors, and ocular-motor functions. Our purpose of study is to investigate blink rate change during the performance of fatiguing prolonged task. It is essential to control several aspects of the task in order to be able to draw conclusions as to how blink rate changes with fatigue. Two factors controlled in our experiment were visual difficulty and required performance rate during a prolonged reading-aloud task.

Demand for high performance. Wood and Bitterman (1950) observed blink rates with the same visual input but under two different demands of performance conditions --- fast reading with maximal effort and slow reading with minimal effort, and reported a significant decrease in blink rate during fast reading as compared with slow reading. They suggested that blink rate might be inversely correlated to performance. A visual tracking experiment by Poulton anf Gregory (1952) demonstrated that blink rate was significantly reduced during tracking as compared with rest periods.

Perceptual difficulty. Different types of tasks result in different blink rates. In a visual tracking experiment, Drew (1951) reported that the blink rate was significantly reduced as the track increased in difficulty from a straight line to an oscillating curve. During the tachistoscopic presentation of gratings at various spatial frequencies and levels of illumination, Mecacci and Pasquali (1980) observed longer latencies of evoked potentials and longer inhibitions of eye blinks at high spatial frequencies and low levels of illuminance. In the reading task experiment, many researchers tried to investigate the effect of level of illumination, presence of glare, and size of type as experimental variables to vary the difficulty of the visual conditions of the task. Some earlier investigations (Lukiesh & Moss, 1937) showed that blink ratte was inversely related to the ease of visual condition under which visual work is performed.

Onthe other hand, certain experimental results (Bitterman, 1945; Bitterman & Soloway, 1946) asserted that blink rate, when it was averaged out over many subjects, tended to be slightly higher under the preferred visual condition. However, in either case, the difference was not significant since individual change of blink rate showed much greater variation with time.

Perceptual attention and cognitive effort. Both perceptual attention and cognitive effort are highly abstract concepts, difficult to quantify; however, it is generally accepted that blink rate is inversely related to them. Kennard and Glaser (1964) reported blinking was inhibited during observation of a light spot and occurred on relaxation or withdrawal of attention. Holland and Tarlow (1975) noted blink rate to be significantly reduced during a nonvisual cognitive backward counting task. In a nonvisual task of recalling numbers with various digit spans, Holland and Tarlow (1972) reported that as the mental load increased with longer digit span, the blink rate for correct trials tended to be lowered. The average blink rate for correct trials were observed significantly lower than for incorrect trials. Other results (Wood & Hasset, 1983; Edwards, 1983) suggested that as cognitive difficulty increased with longer digit numbers, gross blink rate was rather raised.

On the other hand, during conversation (Hall, 1945; Kanfer, 1960) and during slow mental arithmetic with verbalization (Crammon & Schuri, 1980; Schuri & Crammon, 1981), significant increase in blink rate has been reported.

METHODS

Blink recording apparatus. A television eye tracker /pupillometer was used to measure the number of eyeblinks. In preliminary experiments, three outputs --- pupil area, horizontal eye position, and vertical eye position --- were all examined on a chart recorder, and vertical eye position data were stored in LSI 11/23 computer for further analysis. A head rest maintained the enlarged pupil image on the TV monitor screen. The vertical eye position output measured eyeblink reliably and permitted an accurate count of the number of normal full blinks (Figure 1). A simple blinks analysis computer program verified the occurrence of blink and marked these locations with triangles below the time functions of the blink; partial blinks were discarded because of difficulty in distinguishing them from instrumentation glitches (noise).

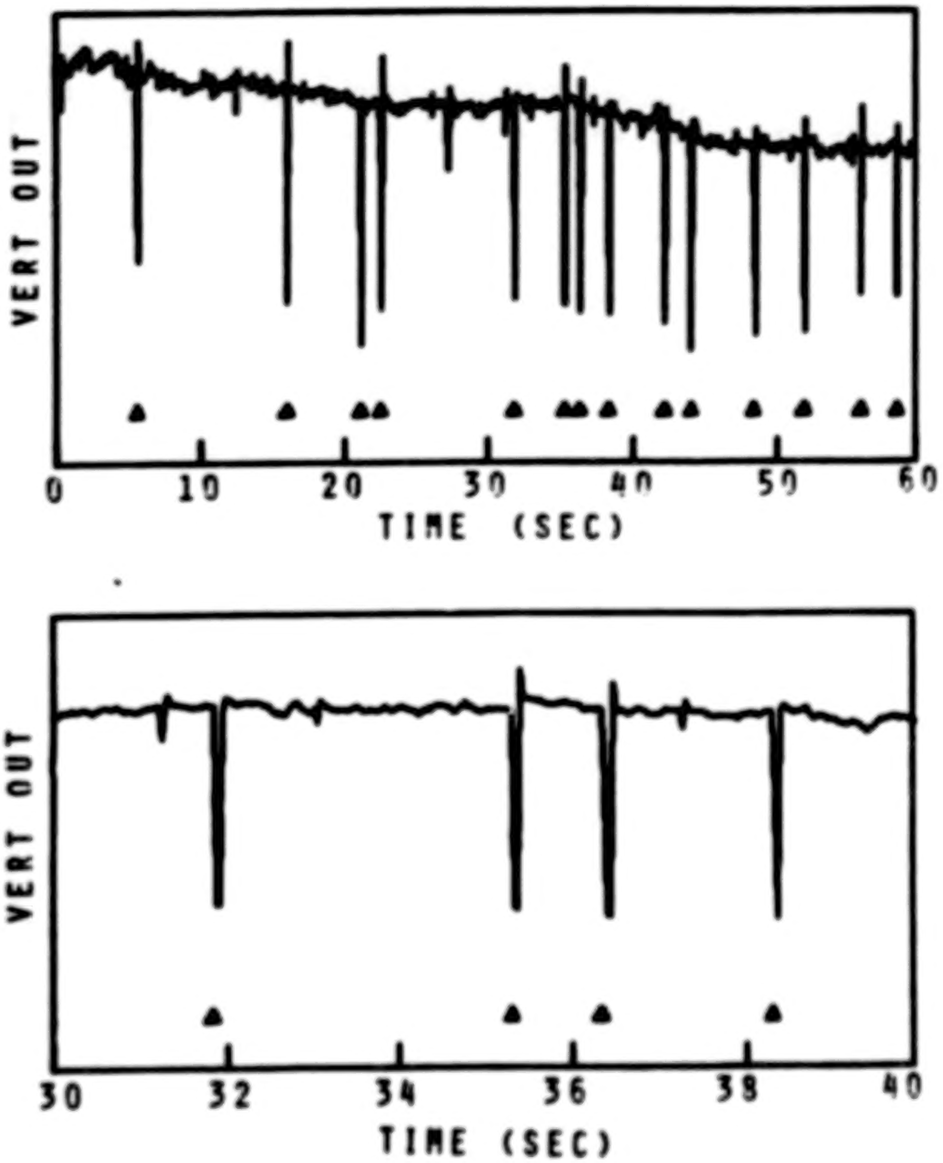


Figure 1 Time recordings of blink. Obtained with TV pupillometer/eye tracker from vertical eye position output channel. Spikes recognized as full blink by computer analysis program are marked with triangles. One minute recording with fourteen blinks (upper graph); a ten second portion on expanded time scale (lower graph).

Reading task presentation. Subjects were seated about 50 centimeters in front of a video monitor screen on which reading texts were displayed. A closed-circuit television system was used to implement blurry as well as clear renditions of text. Considerable defocusing of the video camera produced moderate degree of blurring with luminance and contrast ratios significantly different from clear text (Table I). Fifty sheets were photocopied from a simple high school psychology text and used as the reading material.

Table I

Photometric quantities of the video display screen for the clear and blurry text measured with a Photoresearch Spectra Prichard Model 1980A Photometer

Photometric quantities	clear text	blurry text
Luminance (cd/sqm)		
background (white)	73	66
letter (black)	6	8
overall	49	45
Contrast		
luminance ratio	12.2	8.25
luminance difference	67	58
luminance contrast	0.92	0.88
Size of letter 'I' (min of arc)		
height	30	32
width	4	6

Reading task procedure. One complete experiment of the reading aloud task comprised 9 or 12 sessions, each session lasting 5 minutes. Between sessions, rests were given to subjects for about 1 minute. For sessions 1, 4, 7 and 10, subjects were instructed to read the clear text displayed on the screen aloud and rapidly. For sessions 2, 5, 8 and 11, subjects were instructed to read aloud the blurry text as fast as possible in spite of the marked blur. For sessions 3, 6, 9 and 12, subjects read the clear text aloud but, this time, were instructed to read slowly.

Pseudo-reading task presentation. A "pseudo-reading" tracking target was presented to subjects to produce reading-like patterns. The LSI 11/23 computer generated a seven step staircase waveform and, through the analog-to-digital converter, displayed a point target on the screen. The target repeatedly jumped from left to right in seven steps; each position of the point target required a fixation point as during reading, as well as a return sweep back to the beginning of a new "pseudo-reading" line.

Pseudo-reading task procedure. Subjects were instructed to track the 'pseudo-reading point target displayed on the screen. One complete experiment consisted of eight two-minute sessions with different stimulus rates in the sequence of 0.2, 0.4, 0.5, 1, 2, 3, 4 and 5 stimuli/sec. Between sessions, subjects were given 10 to 50 seconds rest-breaks depending upon their fatigue state. Blink rate was measured both during breaks and during the performance of task.

Combined reading and pseudo-reading task procedure. Reading and pseudo-reading task (at 3 steps/sec or equivalently 25 lines/min) were combined together. Five different operating conditions of 4-minute each were rendered, with 1 minute break between them. The five conditions were :- sequentially, clear-fast, clear-slow, blurry-fast, pseudo readings, and rest. Blink rates during 1-minute rest breaks were also measured.

Subjects Three adults (1 emmetropic male, 1 myopic male and 1 strongly myopic female, all wearing their corrections) participated in this experiment; two repeated the experiment once more on different days to examine within individual consistency of data. Two subjects (1 emmetropic male and 1 myopic male with correction) particpated in the pseudo-reading experiment. One myopic male with correction participated in the combined reading and pseudo-reading experiment.

RESULTS

Blink rate. Different conditions of the perceptual and performance aspects of the task were presented in serial order for five minute sessions each, and the serial order repeated until 12 sessions or approximately 1 hour of reading aloud had elapsed. About 30 seconds to 1 minutes pause occurred between sessions to allow for change of focusing of the TV camera between clear text and blurry text conditions and for change of performance instruction to read aloud fast or slow. Minute-to-minute variations in blink rate (Figure 2) were observed but these variations were much smaller than the marked increase in blink rate (33.4 blinks/min on average) for the clear text slow reading state, as compared with the other two conditions --- blurry text (18.5 blinks/min) and clear text fast reading (17.8 blinks/min): --- not a significant difference in blink rate between these two conditions. These changes continued throughout the more than 1 hour of the experimental run without having seen any significant change due to fatigue. This is so, even though all subjects reported considerable subjective fatigue during the last half hours of the experiments; indeed, one subject stopped at the ninth session. The average blink rates over each of complete 5 minute session again showed clear increase in blink rate for clear text slow reading vis-a-vis over the other two conditions (Figure 3 and Table II). Again, during this fatiguing experiment, no overall increase or decrease in blink rate with time was seen.

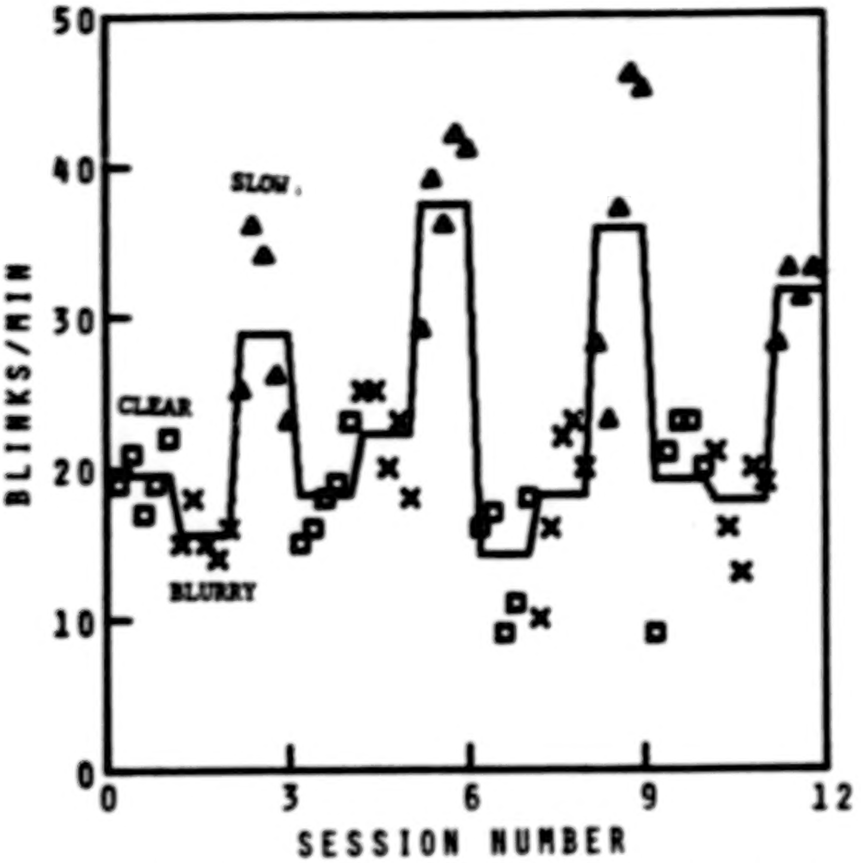


Figure 2 Minute-to-minute variation of blink rate. Rate is number of blink occurrences for each minute (ordinate). Solid line represents mean blink rate for each five-minute reading session (abscissa); sessions were alternated between clear text, fast reading (squares), blurry text, attempted-fast reading (x's), and clear text, slow reading (triangles).

Reading-aloud Performance. The two conditions requiring or producing similar slow blink rates could be easily separated using performance ratings of the simplest kind --- number of lines read aloud per minute (Figure 4). Blurry text fast reading demonstrated low speed reading performance rate (12.2 lines/min on average) due to severe blur (Table III) as compared with clear text fast reading (17.7 lines/min), though the subject tried to read aloud as fast as possible in both conditions. In fact, the performance rate with blurry text was as low as in slow reading with clear text (12.3 lines/min).

Pseudo-Reading. An additional experiment was performed comparing blinking during rest periods with blinking while a tracking a point target that jumped rightward in a seven step staircase and then made a "return sweep" back to the beginning of a new line (Figure 5). The stimuli required through session with

ORIGINAL PAGE IS
OF POOR QUALITY

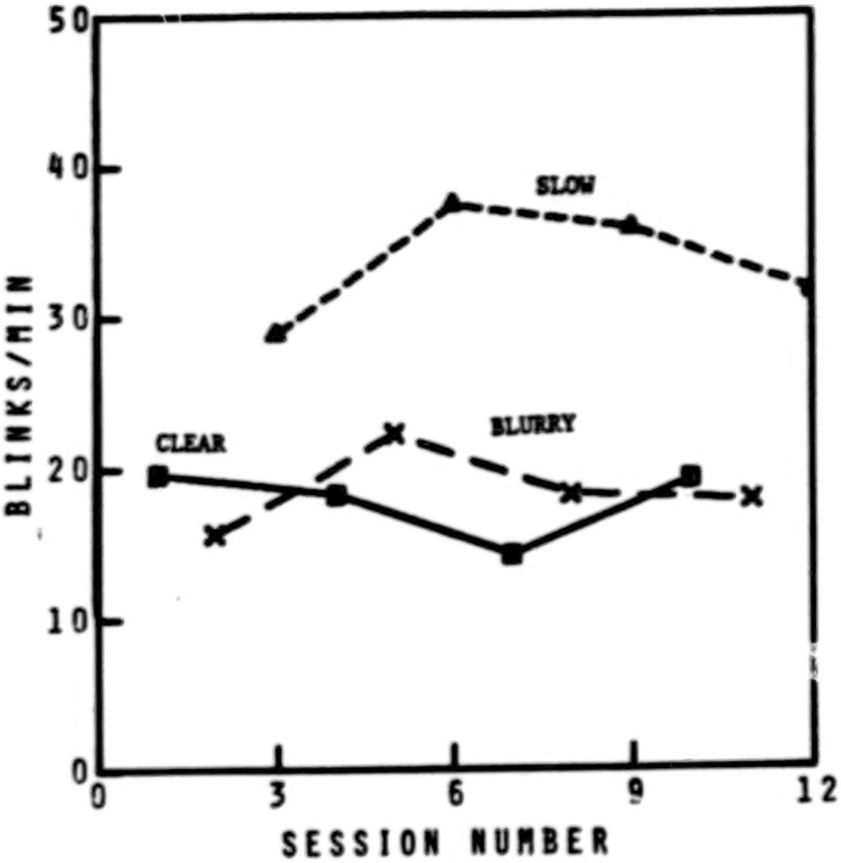


Figure 3 Blink rates during reading aloud task. Average blink rate (ordinate) for each 5-minute reading session (abscissa) with 30 seconds to 1 minute pauses between sessions. Sessions 1, 4, 7 and 10 are clear text, fast reading (squares), sessions 2, 5, 8 and 11 blurry text, attempted-fast reading (x's), and sessions 3, 6, 9 and 12 clear text but with instructed slow reading (triangles). Same experiment as in Figure 2.

increasing rate from 0.2 to 5 stimuli per second. Slower stimuli rates required vigilance in the part of subjects; faster stimuli rates required faster rates of saccadic eye movements. This task was rather effective in reducing blink rate, only 8.1 blinks/min. Conversely, the rest breaks provided an opportunity for the highest blink rates we observed in our experiments, 58 blinks/min. Again, fatigue was not a feature of this approximately twenty minutes long experiment.

Combined reading and pseudo-reading. Two runs of combined five operating conditions experiment were performed. Blink rates under pseudo, clear text fast, and blurry text attempted-fast reading conditions were observed significantly low as compared with clear text slow reading or rest conditions (Table IV). Again, no chanhe with fatigue was seen.

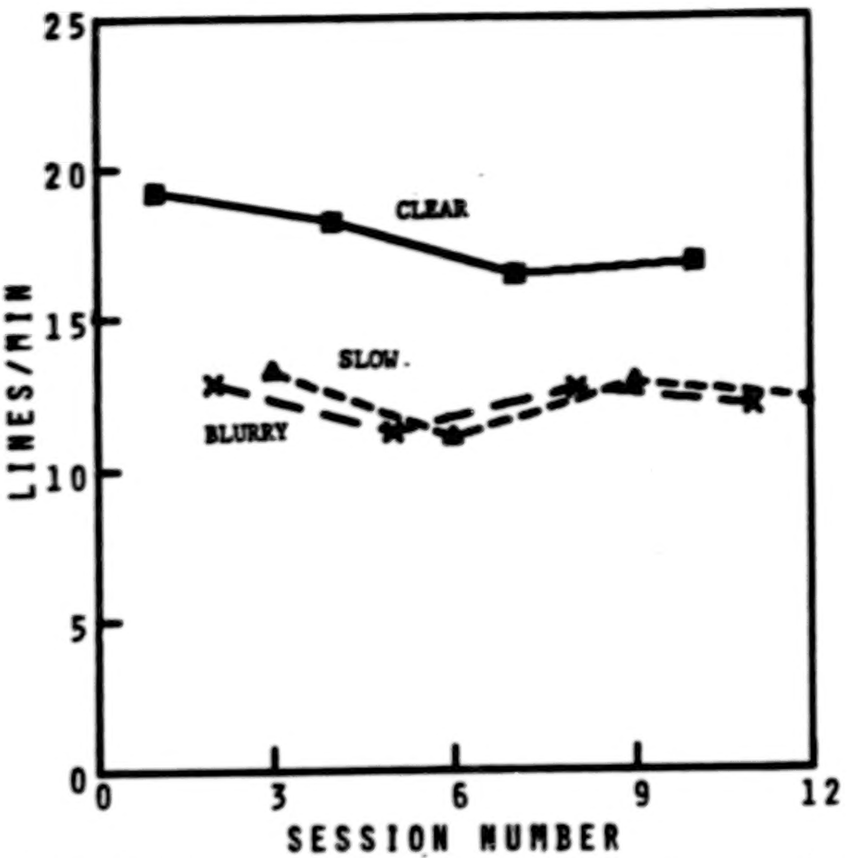


Figure 4 Performance during reading-aloud task. Average number of lines read (ordinate) for each session (abscissa). Same experiment, symbols and abscissa as in Figures 2 and 3.

Variablity. Our three subjects showed quite different mean blink rate under similar experimental condition. For example, 20.7 blinks/minute for subject LA, 34.2 for subject WS, and 10.9 for WZ (Table II, clear text slow reading); others have reported similar inter-subject variation. However, all subjects showed consistent and large reduction in blink rate when blurred text or fast reading conditions existed. Variation between subjects in repeated session was less; even the minute-to-minute variation (Fig. 2) were not so large as to obscure these main effects. Interestingly, variation in mean performance rates (Table III) was less than in mean blink rates.

DISCUSSION

No change in blink rate with fatigue is observed. The task did produce fatigue; our subjects reported complaints of eye strain, backache and neckache. However, they maintained performance rate at approximately constant values, dependent upon operating conditions, throughout one-half or two hour sessions. We conclude that blink rate is not an adequate measure for the detection or assessment of fatigue.

Large changes in blink rate occur dependent upon operating

Table II

Mean blink rates (blinks/min) for three reading conditions

Subjects	clear text fast	blurry text fast	clear text slow
LA#1	8.6	8.0	21.7
LA#2	9.1	11.8	19.6
WS#1	16.5	20.0	35.0
WS#2	17.8	18.5	33.4
WZ#1	2.9	1.6	10.9
Total mean	11.0	12.0	24.1

Table III

Mean performance rates (lines/min) for three reading conditions

Subjects	clear text fast	blurry text fast	clear text slow
LA#1	22.0	13.4	13.5
LA#2	22.3	14.8	13.5
WS#1	18.6	11.2	11.3
WS#2	17.7	12.2	12.3
WZ#1	18.8	12.6	14.3
Total mean	19.9	12.8	12.9

conditions. Rest periods and use of clear text with slow reading aloud as the instructed and monitored task permitted high blink rates. Increased visual difficulty or demand for high performance (fast reading or pseudo-reading) lowered blink rate. These findings support a number of earlier experiments reported (see Introduction) but contrary reports also exist.

As Lukiesh and Moss (1947) suggested, differences in operating conditions may underlie some of the confusion in the literature. For example, if a subject becomes fatigued during the course of an experiment and reduces his task performance (perhaps such changes are also unmonitored by the experimenter), our result would predict an increase in blink rate. This blink rate increase would not be a direct consequence of fatigue on blink rate, but rather an indirect effect due to the strong effect of decreased performance in increasing blink rate.

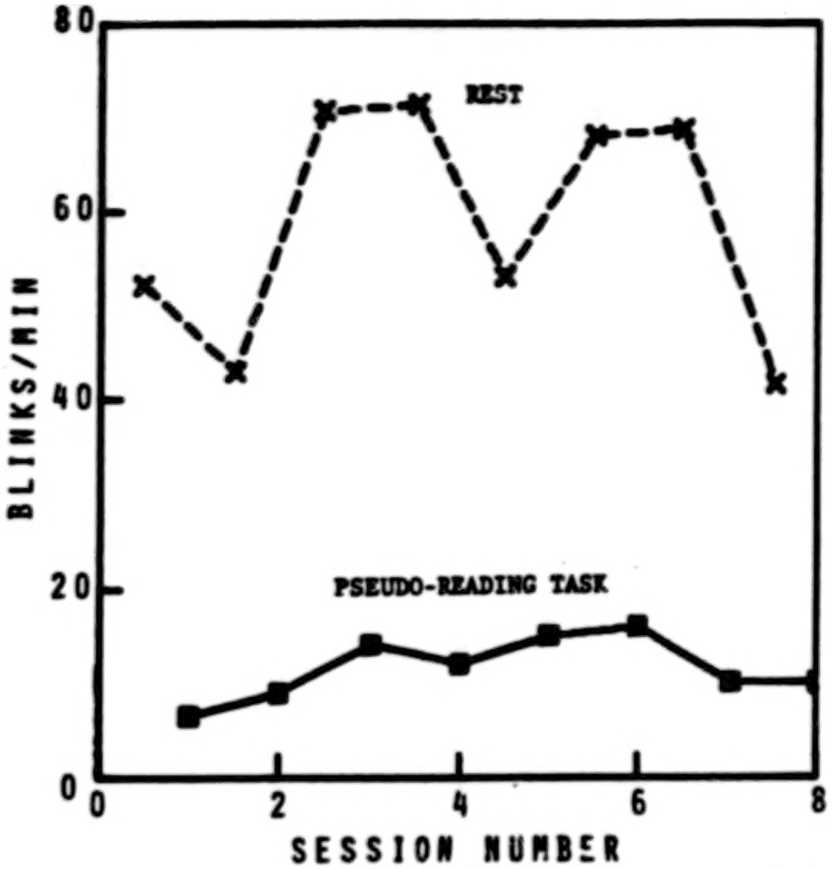


Figure 5 Blink rates during pseudo-reading task. Ordinate: blink rates for pseudo-reading task (squares), and for rest intervals (x's). Abscissa: pseudo-reading session number; each session was 2 minutes long. Step stimuli sequenced through sessions with increasing rate; this had only a minor effect compared to consistent low rate of blinking, 8.1 blinks/min. Rest intervals between sessions were varied from 10 to 60 seconds depending upon subject condition of fatigue; blink rate averaged 58 blinks/min.

ACKNOWLEDGEMENT

We are pleased to acknowledge partial support from NCC2-86 Cooperative Agreement, NASA Ames Research Center, and from the NIH Training Grant in Systems and Integrative Biology. We thank Prof. Victor Gaberseck from University of Paris for his earlier suggestiive blink experiments during his stay in University of California at Berkeley, in the Spring of 1983. We also thank Dr. Dan Greenhouse for quantitative photometric measurement with Prichard photometer, and Dr. Christof Krisher from Julich, Germany for helpful discussion about blurring. We feel very grateful to Dr. Venki Narayan for stimulating discussion on fatigue effect of video display terminal.

Table IV

Relation of visual difficulty and performance rate to blink rate

Conditions	visual difficulty	performance rate	blink rate (blinks/min)
rest periods	Nil	Nil	High (27.0)
clear text slow reading	Low	Low	High (25.6)
blurry text fast reading	High	Low	Low (15.4)
clear text fast reading	Low	High	Low (12.3)
Pseudo-reading	Low	High	Low (14.9)

REFERENCES

Bitterman, M.E. Heart rate and frequency of blinking as indicies of visual efficiency. Journal of Experimental Psychology, 1945, 35, 279-292.

Bitterman, M.E. & Soloway, E. Frequency of blinking as a measure of visuial efficiency: some methodological considerations. American Journal of Psychology, 1946, 59, 676-681.

Carmichael, L. & Dearborn, W.F. Reading and visual fatigue. 1947.

Crammon, D.V. & Schuri, U. Blink frequency and speech motor activity. Neuropsychologia, 1980, 18, 603-606.

Drew, G.C. Variations in reflex blink rate during visual motor tasks. Quartery Journal of Experimental Psychology, 1951, 3, 73-81.

Edwards, J.A. Eyeblink rate as a measure of cognitive processing effort. Ph.D. Dissertation, University of California at Berkeley, 1983.

Florek, H. Spontaneous palpebral reaction in a prolonged visual task. Studia Psychologica, 1972, 14, 313-315.

Hall, A. The origin and purpose of blinking. British Journal of Opthalmology, 1945, 29, 445-467.

Hoffman, A.C. Eye movements during prolonged reading. Journal of Experimantal Psychology, 1946, 36, 95-118.

Holland, M.K. & Tarlow, G. Blinking and mental Load. Psycological Reports, 1972, 31, 119-127.

Holland, M.K. & Tarlow, G. Blinking and Thinking. Perceptual and Motor Skills, 1975, 41, 403-406.

Kanfer, F.H. Verbal rate, eyeblink and content in structured psychiatric interviews. Journal of Abnormal and Social Psychology, 1960, 61, 341-347.

Kennard, D.W. & Glaser, G.H. An analysis of eyelid movements, Journal of Nervous and Mental disease, 1964, 139, 31-48.

Lukiesh, M., Guth, S.K. & Eastman, A.A. The blink rate and ease of seeing. Illumination Engineering, 1947, 42, 584-588.

Lukiesh, M. & Moss, F.K. Reading as a visual task. New York: Van Nostrand Co., 1942.

Mecacci, L. Eyeblink, evoked potentials, and visual attention. Perceptual and Motor Skills, 1980, 51, 891-895.

Ponder, E. & Kennedy, W.P. On the act of blinking. Quarterly Journal of Experimental Physiology, 1927, 18, 89-110.

Poulton, E.G. & Gregory, R.L. Blinking during visual tracking. Quarterly Journal of Experimental Psychology, 1952, 4, 57-65.

Schuri, U. & Crammon, D.V. Heart rate and blink rate responses during mental arithmetic with and without continuous verbalization of results, Psycophysiology, 1981, 18, 650-653.

Tinker, M.A. Reliability of blinking frequency employed as a measure of readability. Journal of Experimental Psychology, 1945, 35, 418-424.

Wood, C.L. & Bitterman, M.E. Blinking as a measure of effort in visual work. American Journal of Psychology, 1950, 63, 584-588.

Wood, J.D. & Hasset, J. Eyeblinking during problem solving: the effect of problem difficulty and internally vs externally directed attention. Psychophysiology, 1983, 20, 18-20.

N85
14565

UNCLAS

N85 14565 D1 30

PERFORMANCE MEASURES FOR AIRCRAFT LANDINGS AS A FUNCTION OF AIRCRAFT DYNAMICS

Edward M. Connelly

Performance Measurement Associates, Inc.
1909 Hull Road
Vienna, VA 22180

ABSTRACT

A theory of performance measurement for operator controlled systems is presented. The theory permits snythesis of a system performance measure which scores performance on successive data samples based on the impact of the sampled performance on the overall summary of performance. Since performance is measured and evaluated on each sample, the dynamics of the controlled element, i.e., the aircraft, are effectively removed from the measurement eventhough the pilot (operator) continues to control the aircraft. While the theory directly applies to problems where the performance limiting factors are known, the method has been extended to apply to problems where the performance limiting factors are not known explicitly, but are known to be implicit in the performance data.

This paper documents the development of measures for aircraft carrier landings for the glide path and angle of attack control channels. While developing the performance measures, the measures used previously were evaluated and were found to lack the necessary discrimination capability. The previously used measure, the RMS of deviations from the glide path, can, for instance, provide identical scores for both satisfactory and unsatisfactory flight paths.

Two types of performance measures developed for aircraft landings are described. Also, an argument is given for the need to test performance measures prior to their use. A suggested test of the measures is offered. Examples of two common measures RMS and "time on target" are identified as having questionable discrimination capabilities. Finally a procedure for synthesizing rather than simply selecting a measure is outlined and illustrated with an example.

INTRODUCTION

A theoretical model based on control theory has been developed to identify the factors that should be determined in both theoretical and empirically-based performance measures. Use of the theory leads to development of comprehensive and sensitive measures.

The theory of performance measurement introduced by Connelly & Schuler (1969) is used here to develop a measurement of the overall task performance in terms of the individual subtask performance effects. This theory was first applied to flight control problems in which the factors limiting performance originated in the hardware and were known. It was recently extended (Connelly, Comeau, & Steinheiser, 1981) to permit its application to team-computer systems where the factors limiting performance are not always known explicitly, but are known to exist.

Since, in many human performance problems, the factors limiting performance are not always explicitly known, demonstrations of task performance at various performance levels that exhibit the effects of those limiting factors must be used to develop the performance measures. This empirically based method for developing measures is described by Connelly, Bourne, Loental, & Knoop (1974) and is the foundation of the Measurement and Analysis of Performance (MAP) Processor. MAP extracts information from the performance demonstration data and then constructs the performance measure. Development of the performance measurement theory, a description of MAP, and applications are given in Connelly (1981).

Types of Performance Measures

Summary Performance Measures. A summary performance measure (SUMPM) is a set of rules for scoring each task demonstration. A SUMPM provides measurement only of the total task performance, and, as a result, the complete information required for a SUMPM is not available until the demonstration has been completed. This property is a fundamental limitation of all SUMPM's.

Typically, SUMPM's are first formulated subjectively, and reflect the judgment of an individual or group concerning the objective of the task and the factors believed to be important in scoring demonstrations. These factors may involve, for example, statements about certain desired terminal and safety conditions that must be satisfied during the task demonstration. But whatever the factors are, the subjective form of the SUMPM must then be converted into a quantitative form in which specific rules determine the SUMPM value from the demonstration.

The central issue regarding the selection of a SUMPM is its ability to discriminate performance of task demonstrations. A SUMPM is said to be acceptable only if it scores (or at least orders) performance demonstrations the same way the investigator (or group accepted as subject matter experts) would score or order performances. An unacceptable SUMPM would score performances that are actually poor at a higher value or performances that are actually superior at a lower value. Consequently, some superior performances demonstrations might be rated as fair or poor and vice versa.

The consequence of using a measure whose acceptability is unknown, as might be the case if a SUMPM is simply selected without further analysis, to evaluate performance in experiments supporting decisions regarding such things as training courses, and equipment designs is to transfer the uncertainty of the measures' discrimination capability to an uncertainty of the decision. Yet, in spite of the serious consequences of using a measure with an unknown acceptability, few studies report results of measure evaluation or the measure snythesis procedures used to control the measures' acceptability. In these cases one must suspect that the measure was simply "selected" and its acceptability is, in fact, unknown.

Two SUMPM measures frequently used in aircraft flight control and weapon aiming tasks (as well as many other tasks) that have been found to have poor discrimination capability are the root-mean-square (RMS) and time-on-target (TOT) measures. Early suspensions concerning these measures were reinforced when performance discrimination failed to show significance as reported in Connelly, Schuler, Knoop (1971) and again in Connelly, Bouren, Loental, Knoop (1974). Poulton (1974) also reports the lack of discrimination capability of TOT. Further, Connelly, Shipley (1981) show that even a "super" TOT where both error and error rate must be near zero (within a small tolerance) does not have acceptable performance discrimination capabilities. Regarding the RMS error, Shipley (1983) shows various charts of aircraft altitude control which include converging, diverging, exponential, and oscillatory responses; but, all have the same RMS score!

The solution to the performance measurement problem for flight control is conceptually simple. Typically for flight control problems an isolated reference (i.e., the desired) path is established such as constant altitude or a glide path and all other trajectories are scored based on their distance from the reference - according to the RMS function. By specifying a reference path and a scoring rule for trajectories off-the-reference-path, reference trajectories not on the reference path are implied, but do exist even though they are not directly specified. These off-the-reference-path trajectories, which may not be known to the measure user, are the optimal trajectories, i.e., those that minimize the RMS measure. It is quite possible that the optimal control responses required to produce the "optimal" trajectories, which can be "bang-bang" control responses (control device fully on or off), may not even be desirable control strategies. A solution to this problem is to specify reference trajectories everywhere of interest in the problem space rather then only specifying an isolated reference path and an error criterion. Reference trajectories can be conveniently specified by the differential equation that produces them. This paper describes the results of such a specification in an aircraft landing problem. A more complete description of the method which includes a methodology for snythesizing acceptable SUMPMs, using the MAP Processor, is found in Connelly (1981).

System Performance Measures. System performance studies conducted to evaluate such things as unit readiness level, or the effects of different types of equipment, training, procedures must cope with the problem of varying effect on performance of these factors at various parts of a task. If, for instance, the use of new equipment affects performance in different ways or to a different degree at various parts of a task, then it is difficult to evaluate that affect using the SUMPM. This is because the SUMPM reflects performance of the parts of the task affected by the new equipment plus the other portions of the task not affected by the new equipment, but affected by all the other performance influencing factors. The other performance influencing factors are considered as "noise" tending to obscure the relationship of interest. If a measure could be developed that permits assessment of the affect of performance on each part of a task on total task performance, then task performance can be evaluated by observing performance on only critical portions of the task. This would greatly improve the measure sensitivity.

System performance measures (SYSPM) reveal the effect of performance of a constituent part of a task on summary performance and as a result, provides sensitive performance discrimination. System performance measurement theory, which was developed first by Connelly, et al. (1969) and later extended by Connelly, Zeskind, & Chubb (1977), recognizes that performance is limited both by machine factors and by human factors. Recognition that such limiting factors exist, whether or not they are explicitly known, leads to a measurement equation that permits evaluation of the effect of either instantaneous or of interval performance on the performance of the entire task.

A second issue regarding system performance measures results from the ability of SYSPM to score performance at closely spaced data samples. For instance, some data collection systems sample at rates of 10 to 100 samples per second and control performance can be scored for each sample, i.e., the SYSPM removes the dynamic lag effects of the aircraft from the score calculation. But performance scores from successive samples can be correlated due to the dynamic response of the pilot. Consequently, the measurement issue is: what time difference between data samples permits independent evaluation of successive samples. If the time between samples is large, each can be considered as representing an independent control problem to the pilot. If, for example, a task lasts 60 seconds and the task can be partitioned into independent parts every 6 seconds, then there would be 10 independent control problems presented to the pilot. Instead of 1 performance score value, there would be 10 repetitions scores. As the partition interval is decreased, there is an increase in the number of repetitions (N) (assuming that all the control problems are considered to be the same). However, the partition interval reduction is limited - due to the dynamic processing and response time of the pilot. Clearly intervals less than expected human response times (from 0.05 to 0.10 seconds) may exhibit correlated responses. In order to investigate this issue, a correlation analysis of successive performance measurement samples for various partition intervals was performed. The results are presented in a subsequent section.

Method of Approach

The method employed here can be applied where performance demonstration data are available but the demonstrations are not scored or ordered according to performance. The method uses demonstration data to develop a model of the system (aircraft) dynamics and also performance criteria. To accomplish this, demonstration data is analyzed to develop a set of approximate aircraft equations and representative pilot control policies. These equations and policies are used to construct a measure which will indicate the convergence of at least some performance demonstrations. The measure is then tested and applied to all performance demonstrations so that all demonstrations are scored consistently.

The specific approach was to construct a second order performance model which measures performance according to how well the pilot controls the second error derivative given the error and its first derivative. For instance, in logitudinal glide path control, the second derivative of the glide path error is evaluated as a function of the glide path error and its first derivative.

In general, a model of any order desired for a specific problem can be constructed. If it is known, for instance, that the pilot's (operator's) controls directly affect the first derivative of the error, then a first order model should be used. The objective is to use a model of an order that permits evaluation of a derivative that can be or is rapidly adjusted by the pilot (operator). The ideal model would permit evaluation of the control element (throttle, control stick) as a function of system state. However, as will be seen, it is desirable for computational simplicity to evaluate performance of a variable which is dynamically "close" to the pilot's control elements, i.e., a variable that can be rapidly modified by the pilot.

Specific Method of Approach. Take the set of equations as:

$$\dot{X}_1 = X_2 \quad (1)$$

$$\dot{X}_2 = -aX_1 - bX_2 + U \quad (2)$$

where

X_1, X_2 are state variables and U is a control variable. The summary performance measure function is taken as:

$$I = \int_0^{t_1} F(X_1, X_2, U)\, dt \quad (3)$$

where

$$F = A_1X_1^2 + A_2X_1X_2 + A_3X_2^2 + A_4U^2 \qquad (4)$$

and $F > 0$ except at the origin. Conditions insuring that $F > 0$ are that $A_1A_4 > 0$ and the quadratic has imaginery roots. This requires that

$$A_2^2 < 4A_1A_3 \qquad (5)$$

Optimal control theory is used to find the optimal control law and the system performance measure which has the form

$$SYSPM = A_4 (U-U^*)^2$$

where U^* is the optimal control law

$$U^* = \frac{1}{2A_4} (B_1X_1 + 2B_2X_2)$$

and B_1 and B_2 are functions of A_1, A_2, A_3 and A_4. A detailed solution for SYSPM is given in Connelly (1982).

RESULTS

Results concerning the effect of the command and conventional displays are presented in Westra (1982). Results regarding the application of the measures to the landing problem are given in Connelly (1983). Development of the system performance measure showed that a measure referencing the system differential equation rather than the isolated glide path provides greatly improved performance discrimination. With this approach, performance deviations are determined by comparing the observed value of the rate of change of each system variable with a reference differential equation. Consequently, there exists a family of differential (incremental) reference solutions everywhere in the problem space where an acceptable solution is possible. This is in contrast to the commonly used approach where an isolated glide path is the reference and performance deviations are measured as the distance of the aircraft from the glide path.

An auto correlation analysis was performed to determine the correlation of performance scores for data samples shifted T samples from each other. T was varied from 1 to 10. Since samples were taken at a rate of 30 times a second, each shift is equivalent to a time difference of T/30 seconds. It was expected that correlations would be high for small T with a reduction in the correlation coefficient for increasing T. Results showed that the correlation for T=1 was high, being in the order of .95. But correlation values for large T shifts (2 through 10) dropped to a low value - in the order of .005. Such a rapid reduction in correlation coefficient values, with increasing T, cannot logically be attributed to the speed of response of the pilot because it is known that control elements were held constant for longer periods, e.g., in some instances the throttle was maintained at a constant level for several seconds. Thus,there is evidence that the system performance measure, which adjusts the reference control as a function of the state variables, presents independent problems to the pilot at each sample. Additional study is required to investigate this issue since we cannot expect a pilot (or other human operator) to respond to each sample when T is small; but, when T is large each sample is an independent test and each flight, which consists of many samples, will contain many independent tests.

C-5

REFERENCES

Connelly, E. M. Development of system performance measures. Performance Measurement Associates, Vienna, Virginia, 1981.

Connelly, E. M. Performance measures for aircraft carrier landings as a function of aircraft dynamics. (Tech Report NAVTRA-EQUIPCEN 80-C-0132-1). January 1982.

Connelly, E. M. Performance measures for aircraft carrier landings as a function of aircraft dynamics. Paper presented at the Human Factors Conference, Norfolk, VA. 1983.

Connelly, E. M., Bourne, F. J., Loental, D. G., & Knoop, P. A. Computer aided techniques for providing operator performance measures. (Report No. AFHRL-TR-74-87). December 1974).

Connelly, E. M., Comeau, R. F., & Steinheiser, F. Team performance measures for computerized systems. (Final Tech Report for ARI Contract No. MD 903-79-C-0274). 1981.

Connelly, E. M. & Schuler, A. R. A theory adaptive man-machine systems applied to automated training. Paper presented at the International Symposium on Man-Machine Systems, St. John's College, Cambridge, England, September 1969.

Connelly, E. M. & Shipley, B. D. An analytical model for developing objective measures of air crew proficienty with multivariate time sequenced data. (Tech Report on Contract No. MDA 903-80-C-0198). May 1981.

Connelly, E. M., Zeskind, R. M., & Chubb, G. P. Development of a continuous performance measure for manual control. (Final Report on Contract F33615-75-C-5088, AMRL-TR-76-24).April 1977

Poulton, E. C. Tracking skill and manual control. Academic Press: New York, N.Y. 1974.

Shipley, B. D. Maintenance of level flight in a UH1 flight simulator. Army Research Institute, Flight Unit of Ft. Rucker, Ala. (In press) 1983.

Westra, D. P. Simulator design features for carrier landing: II. In-simulator transfer of training. (Tech Report NAVTRAEQUIPCEN 81-C-0105). NTEC 1982.

N85
14566

UNCLAS

D31

N85 14566

Measuring pilot workload in a moving-base simulator: II. Building levels of workload

Barry H. Kantowitz, Sandra G. Hart, Michael R. Bortolussi,
BITS, Inc. NASA-Ames BITS, Inc.
693N 400W Rsch Center
W. Lafayette IN Moffett Field CA

Robert J.Shively Susan C. Kantowitz
Dept of Psychology BITS, Inc.
Purdue University
W. Lafayette IN

Studies of pilot behavior in flight simulators often have used a secondary task as an index of workload (e.g.,Kantowitz, Hart, & Bortolussi, 1983; Wierwille & Connor, 1983). It is routine in such studies to regard flying as the primary task and some less complex task as the secondary task. Thus, flying is considered a unitary task much as the secondary task is considered to be a unitary task. While this assumption is quite reasonable for most secondary tasks used to study mental workload in aircraft (Williges and Wierwille, 1979), the treatment of flying a simulator through some carefully crafted flight scenario as a unitary task is less justified. While this is often a necessary simplification that can be easily forgiven since it yields useful information, it should be remembered that flying is a complex task that is likely to have an hierarchical organization. While researchers concerned with training have never forgotten this, researchers who are concenred with evaluating workload with skilled pilots tend to ignore the general complexity of flying and have been content to acknowledge only the general difficulty of a particular flight scenario with little regard to complexities that might be related to the hierarchical structure of the flight task.

The present research is a first step towards acknowledging that total mental workload depends upon the specific nature of the sub-tasks that an aircraft pilot must complete. As a first approximation, we have divided flight tasks into three levels of complexity. The simplest level (called the Base level) requires elementary maneuvers that do not utilize all the degrees of freedom of which an aircraft, or a moving-base simulator, is capable. Examples would be flying at a constant altitude or at a constant heading. The second level (called the Paired level) requires the pilot to simultaneously execute two Base level tasks, for example, flying on a constant heading while also maintaining a constant altitude. The third level (called the Complex level) imposes three simultaneous constraints upon the pilot. An example would be flying at a constant altitude, on a constant heading, and at a constant speed. Further example of Base, Paired, and Complex tasks used in this experiment can be found in Table 1. Note that even the Complex level is relatively

PRECEDING PAGE BLANK NOT FILMED

elementary when compared to the actual demands of flight where other necessary tasks such as navigation and communication must also be performed. This additional complexity is addressed in Experiment II, currently in progress.

Workload is assessed by subjective ratings and by an asynchronous secondary choice-reaction task quite similar to those used by Kantowitz, Hart and Bortolussi (1983). Two general questions are asked. The first involves comparing secondary-task performance under single- and dual-task conditions. Since highly skilled pilots are being tested, one reasonable prediction would be that elementary maneuvers are so automatic and overlearned that they impose no workload on the pilot. Therefore, one would expect no differences between secondary-task performance regardless of whether or not the primary flying task was required. An alternate prediction, based upon the notion that training does not eliminate attentional requirements of flight (Johnson, Haygood & Olson, 1982), would expect faster reaction times and/or fewer errors under single-task conditions. The second general question arises only if the alternate prediction is correct. Given that even these elementary flight tasks create workload, one can then ask if the three different levels of task complexity defined a priori as Base, Paired, and Complex also produce different levels of pilot workload. One might expect that task differences, especially between Base and Paired, are so small that no workload differences should be produced or one might predict that workload should increase as levels go from Base to Complex. And of course, one can always ask the eternal question in workload studies by attempting to relate subjective and objective measures of pilot workload.

METHOD

Pilots

Seven male and five female instrument-rated pilots served as paid participants. Four pilots had a private pilot license, six had commercial licenses, and two had airline transport licenses. Pilots had from 500 to 6000 hours of total flight time (median=1025 hours) and from 30 to 1200 hours of actual instrument time (median=130 hours).

Flight Tasks

Each pilot flew 21 separate flight tasks (Table 1) twice, once with the secondary task and once by itself. Each flight task lasted three minutes. All flight tasks were flown in a Singer/Link GAT-1 instrument trainer with three degrees of freedom. As indicated in Table 1, certain degrees of freedom were frozen for certain flight tasks. This prevented the pilot from attempting to control irrelevant simulator motion. Freezing a task component also froze the corresponding instruments inside the simulator.

TABLE 1

BASE LEVEL-TASK.	PITCH	ROLL	YAW	ALT	ASI
1. FLY HDG 360	F			F	F
2.MAINTAIN 2000FT.		F	F		F
3. "S" TURN	F			F	F
4. CLIMB AT 500FPM		F	F		F
5. DESCEND AT 500FPM		F	F		F
6.MAINTAIN 120KTS.		F	F	F	
PAIRED LEVEL-TASKS					
1. FLY HDG 360,MAINTAIN 2000FT.					F
2. MAINTAIN 2000FT.,"S" TURN					F
3. FLY HDG 360,CLIMB AT 500FPM					F
4. FLY HDG 360,DESCEND AT 500FPM					F
5. "S" TURN,CLINB AT 500FPM					F
6. "S" TURN,DESCEND AT 500FPM					F
7. FLY HDG 360,MAINTAIN 120KTS				F	
8. MAINTAIN 2000FT.,MAINTAIN 120KTS		F	F		
9. "S" TURN, MAINTAIN 120KTS.				F	
COMPLEX LEVEL-TASKS					
1. FLY HDG 360,MAINT 2000FT.,MAINT 120KTS.					
2. FLY HDG 360,DESC. AT 500FPM,MAINT 120KTS.					
3. FLY HDG 360,CLIMB AT 500FPM,MAINT 120KTS.					
4. "S" TURN 360,DESC. AT 500FPM,MAINT 105KTS.					
5. "S" TURN,CLIMB AT 500 FPM,MAINT 105 KTS.					
6. "S" TURN,MAINT 2000FT.,MAINT 120KTS.					

Secondary Task

Three positions of a helicopter trim switch ("coolie-hat" switch) mounted on the left side of the control yoke under the pilot's thumb were used for responses to auditory tones. A low tone (800 Hz) was paired with switch motion to the left, a medium tone (1500 Hz) with a forward switch motion, and a high tone (4000 Hz) with a right switch motion. Tones were 300 msec in duration and approximately 70 dB SPL, presented over headphones. An Apple II computer with a Cyborg Model 91A interface generated tones and recorded reaction time to the nearest millisecond as well as errors. Tones were presented asynchronously--that is, regardless of performance on the flying task--every eight seconds.

Normally, it is prudent to utilize two levels of difficulty in the secondary task (Kantowitz & Knight, 1976) to ensure that data can be theoretically interpreted. However, only one level (3-choice task) was used in this study because an earlier study using much the same secondary task (Kantowitz, Hart & Bortolussi, 1983) found no interaction with two- and four-choice auditory secondary tasks.

Procedure

Each of the 21 flight tasks were flown twice: with and without the secondary RT task. As a single-task control condition, the RT task was performed alone in the GAT cockpit at the end of each flight level. All 3! orders of flight level were used with two subjects randomly assigned to each order. In each block half of the pilots flew the task with tone first (dual-task condition) and the other half flew first without tones (single-task condition).

All pilots were given approximately 30-40 minutes of simulator practice to learn the flight characteristics of the GAT before starting the experiment proper. Practice on the auditory choice-reaction task continued until a criterion of 95% -98% accuracy was achieved.

Immediately after each single-task flight condition, pilots completed bipolar rating scales for ten items. During all simulated flight airspeed, altitude, x-y position and rudder, elevator and aileron control deflection were continuously recorded.

RESULTS

Primary Task Performance

The major concern to be evaluated is a comparison of single- versus dual-task performance for the flying task. The relative

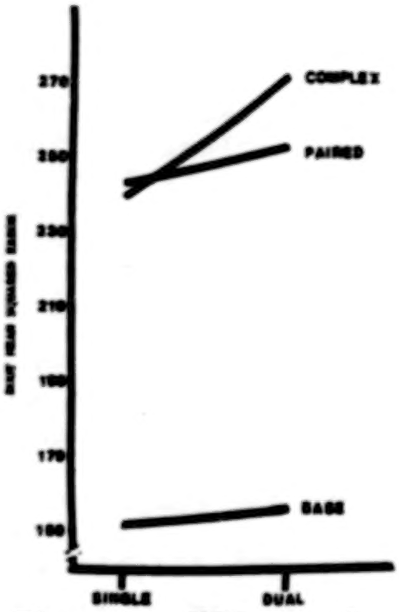


Figure 1. RMSE as a function of single vs. dual task performance.

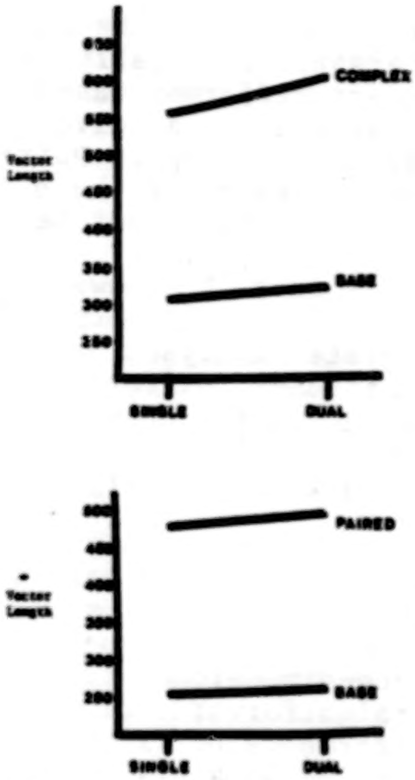


Figure 2. Vector length as a function of base vs. paired and base vs. complex.

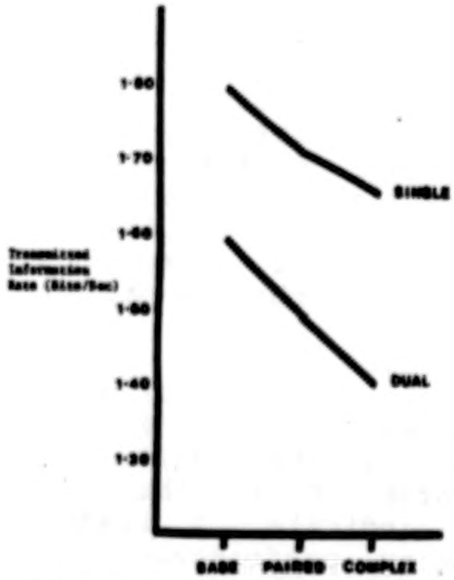


Figure 3. Secondary task performance (bits/sec) as a function of level of the primary task averaged over blocks.

performance for the 21 flight tasks of Table 1 is not of major interest, especially since it is not clear how to directly compare different tasks, e.g., how much rms error in altitude is equivalent to a given rms error in heading? It is, however, possible to compare Paired and Complex tasks with the appropriate Base tasks since here the units are comparable but Paired and Complex tasks cannot be contrasted.

Figure 1 shows rms error for single- versus dual-task performance for each of the three levels of complexity. Three separate analyses of variance (one at each level) revealed no significant differences between flying alone and flying plus responding to tones for the Base level (F(1,11) = 0.04), Paired level (F(1,11) = 0.54), and Complex level (F(1,11) = 0.18). Thus, adding the secondary-tone task did not alter flying performance.

A vector analysis was computed in order to contrast Base versus Paired and Base versus Complex flight performance. This is best illustrated by the Base versus Paired comparison which can be plotted in two-dimensional space but the extension to the three-dimensional space of the Base versus Complex comparison is straightforward. Let us select as an example a comparison of Base performance of flight tasks 1 and 2 in Table 1 with flight task 7 that demands simultaneous performance of tasks 1 and 2. In a two-dimensional space we can plot Base performance with rms error in heading as a point on the abscissa and rms error in altitude as another point on the ordinate. Paired performance can be represented by a single point in this vector space. We then calculate the length of the existing vector representing Paired performance and also the length of the implied vector formed by projecting the two Base points perpendicular to their respective axes until they meet. Note that this implies an equal weighting of the scales shown on the abscissa and ordinate and that such an assumption requires empirical justification which we shall soon provide. Figure 2 shows comparisons based upon vector length. As we would expect from Figure 1, there was no significant effect of single- versus dual-task for either the Base vs. Paired comparison, F(1,384) = .14, or the Base vs. Complex comparison, F(1,240) = 1.03. However, significant effects indicating reliably smaller rms error in the Base condition were obtained for both Base vs. Paired, F(1,384) = 31.63, p<.001, and Base vs. Complex, F(1,240) = 32.55, p<.001, comparisons. No significant interactions were obtained for either comparison.

In order to check the validity of the equal-weighting assumption mentioned above, an additional analysis was performed whereby the length of a vector's projection upon an axis was compared to Base performance on that axis. If performance for the Base condition was worse than the corresponding vector projection, this might indicate a trade-off between task components where outstanding performance on one task component (i.e., performance better than that component performed singly during the Base condition) was achieved at the expense of performance on the remaining vector projection(s). There were 43

ORIGINAL PAGE IS
OF POOR QUALITY

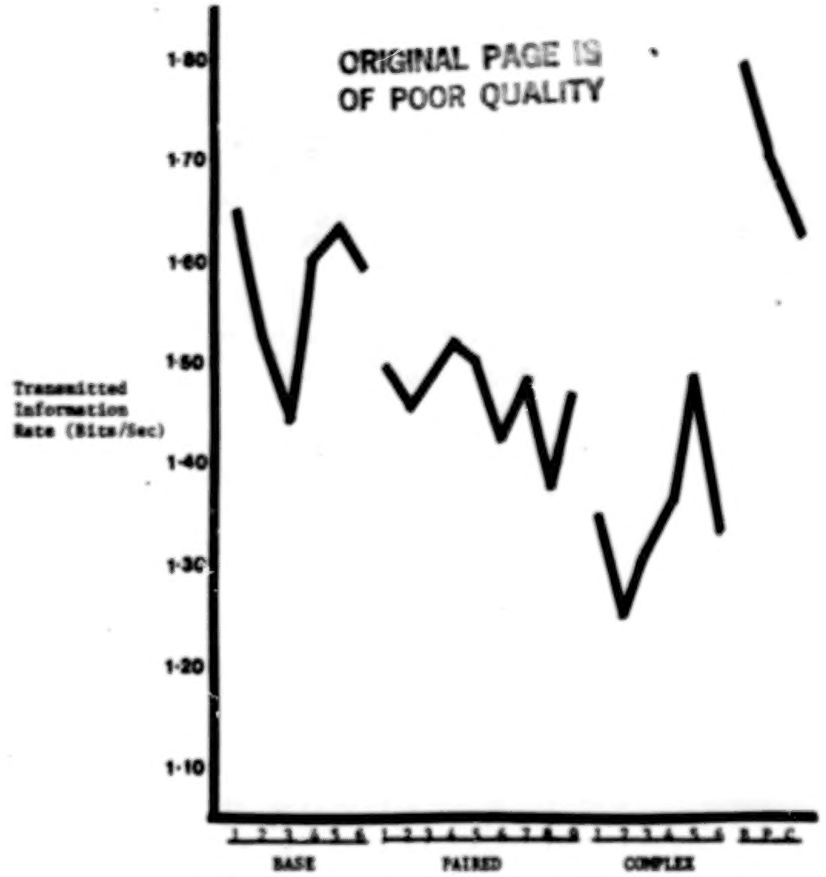


Figure 4. Secondary task performance (bits/sec) as a function of level of the primary task.

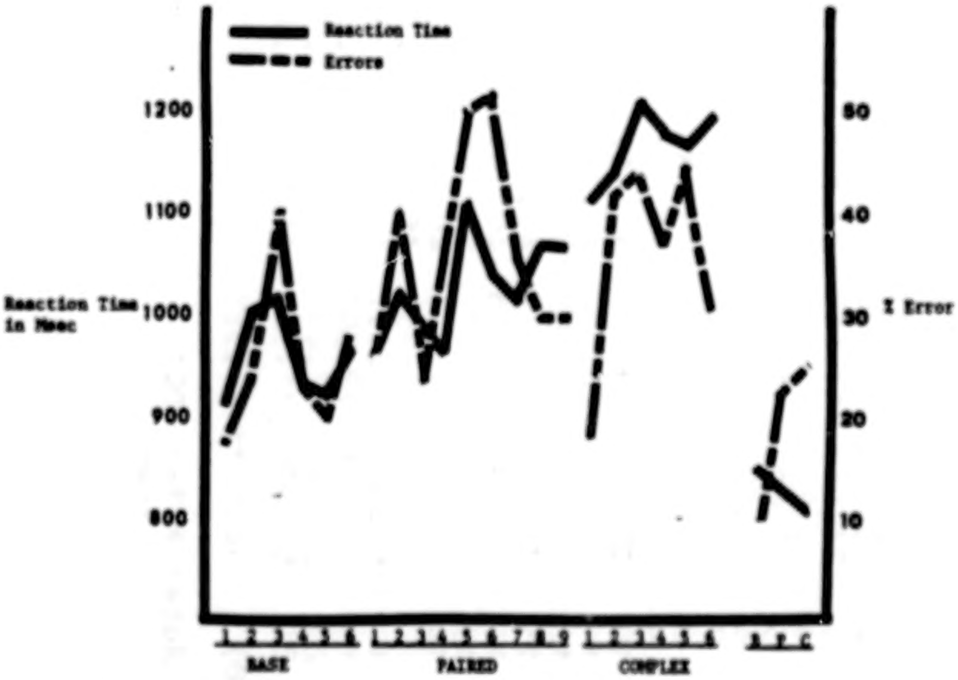


Figure 5. Secondary task performance (Reaction time/Errors) as a function of level of primary task.

ORIGINAL PAGE IS
OF POOR QUALITY

FLIGHT TASK	Task Difficulty	Time Pressure	Performance	Mental Effort	Physical Effort	Frustration	Stress	Fatigue	Activity Type	Overall Workload
1	12.917	8.833	29.917	34.000	13.583	21.917	15.583	32.167	9.250	22.667
2	19.333	13.000	31.250	30.500	16.583	26.667	19.917	29.833	10.000	24.167
3	24.667	14.667	32.583	41.750	25.167	30.583	23.167	33.167	8.000	28.750
4	17.750	10.750	31.833	31.833	17.917	26.000	23.583	35.250	9.833	19.583
5	22.000	13.417	36.667	36.250	18.833	28.000	23.833	35.000	9.833	25.750
6	15.583	9.167	32.333	36.500	12.917	23.417	17.667	35.000	6.833	17.750
7	32.583	17.417	44.083	51.417	24.500	35.417	24.333	37.000	9.333	39.167
8	41.583	24.750	53.333	57.417	25.750	42.583	30.583	41.917	14.000	43.833
9	38.667	19.083	49.333	48.250	20.583	37.500	25.583	42.250	15.583	46.500
10	37.000	19.083	45.917	48.500	25.333	31.417	24.417	45.333	10.500	45.583
11	45.750	20.750	44.250	56.000	24.333	29.583	23.333	43.083	11.583	51.750
12	50.167	19.583	50.833	52.667	23.167	38.667	28.917	48.083	12.667	52.833
13	31.167	12.167	46.750	45.667	21.833	29.917	25.000	47.000	10.667	38.000
14	30.583	13.083	40.167	39.167	20.167	28.667	21.583	39.750	6.667	32.750
15	36.667	13.417	54.583	50.250	18.333	36.750	26.333	41.333	8.250	42.667
16	41.667	15.833	40.167	44.833	23.750	27.167	25.667	39.500	13.750	46.333
17	48.167	16.750	48.083	50.750	32.583	37.333	31.417	46.833	15.167	52.333
18	52.000	17.917	51.167	57.500	25.167	36.750	33.750	49.333	15.583	61.250
19	53.333	19.500	49.917	60.333	32.000	45.167	40.167	51.583	15.250	59.750
20	57.083	22.417	54.333	60.833	35.083	45.333	36.833	52.083	16.417	58.500
21	53.000	21.500	45.833	54.500	30.667	29.417	28.583	50.417	14.667	54.500
F(20,220)	16.1***	1.78*	3.02***	9.06***	3.56***	2.44**	3.62***	2.96**	1.95*	17.1***

Note: ***=p .001
**=p .01
*=p .05

Table 2. Mean Subjective Ratings

TASK	MEAN	df	F
Base tasks 1-6		5,55	2.73*
1	20.4		
2	22.2		
3	25.4		
4	22.3		
5	24.7		
6	20.6		
Paired tasks 7-15		8,88	2.74*
7	29.8		
8	35.9		
9	31.3		
10	31.3		
11	32.1		
12	34.5		
13	29.8		
14	25.8		
15	31.1		
Complex tasks 16-21		5,55	3.54**
16	29.5		
17	35.9		
18	36.2		
19	39.5		
20	41.3		
21	36.4		

Note: **=p .01
*=p .05

Table 3. Weighted Mean Subjective Ratings

possible paired comparisons of this nature for single- and also for dual-task performance. Since there were 12 subjects a total of 1032 data points were examined (43 X 2 X 12). We searched for cells in which at least 9 subjects showed lesser vector projections since this would be a significant number of subjects by sign test. Of the total of 86 cells (43 single- and 43 dual-task) only three cells had 9 such deviant points and no cell had 10 or more deviant points. Hence, we conclude that an equal-weighting assumption is reasonable for these data.

To recapitulate, the tortuous analysis of primary task performance, required since the various rms error scales are not equivalent, showed that Base performance was better than either Paired or Complex performance. This is hardly an astonishng outcome and the detailed vector analysis should not detract from the more important result shown in Figure 1 that addition of a secondary task did not alter primary flight-task performance.

Secondary Task Performance

For each pilot and each flight task, transmitted information (bits/sec) was calculated for the secondary three-choice reaction task. Since this measure takes both speed and accuracy into account, it is the optimal index of secondary-task performance (Kantowitz, Hart, & Bortolussi, 1983). Figure 3 shows that transmitted information was highest for the Base level conditions and declined with higher flight-task levels, F(2,22) = 8.23, p<.001. As was expected, reliably more information was transmitted during the single-task control conditions, F(1,18) = 39.6, p<.001. However, while transmitted information was able to discriminate among levels of flight task, three separate analyses of variance performed within each level (Figure 4) were unable to detect any reliable differences.

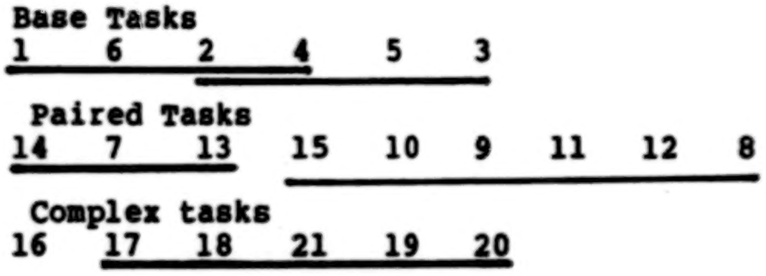
Figure 5 shows the same results as Figure 3, except that reaction time and errors are plotted separately rather than combined as transmitted information. Effects of level were significant for both reaction time, F(2,252) = 33.1, p<.001, and errors, F(2,252) = 4.12, p<.05.

Subjective Ratings

Subjects were asked to rate each of 21 flight tasks using 10 bipolar rating scales. The results of the analyses of variance indicate that all the scales were able to distinguish between at least two of the 21 flight tasks (Table 2.).

Further analysis was done to determine the effect of flight task on rating behavior. The subjects gave a subjective rating of importance to each of the 10 scales. This importance rating was used to weight each subjects summed ratings on all the scales for each of the 21 flight tasks. Analysis of the weighted mean scores over all flight tasks indicates that at least 2 of the 21 flight task means are significantly different, F(20,220)=8.84, p<.001. The flight tasks were divided into 3 categories and separate analysis were calculated on the mean scores in each category. The results indicate that at least 2 of the means for each category differ significantly (Table 3).

To determine which flight task means differed t-tests were calculated on all possible pairs of flighttasks within each category. The significant mean differences are summarized in Table 4.



Base Tasks

1 6 2 4 5 3

Paired Tasks

14 7 13 15 10 9 11 12 8

Complex tasks

16 17 18 21 19 20

Tasks are arranged in increasing mean value for each category. The line indicates those means that do not differ significantly at $p<.05$.

Table 4. Pairs of Flight Tasks

DISCUSSION

Results clearly showed that even the most elementary flying tasks (Base) produced measurable pilot workload using the objective secondary-task technique. Furthermore, as the flying tasks were made more complicated, progressing to Paired and Complex tasks, workload increased even more. These findings are impressive confirmation of the utility of the asynchronous choice-reaction secondary task used by Kantowitz, Hart and Bortolussi (1983). Primary task performance was unaffected by the addition of the secondary tone-task while transmitted information decreased with flight-task complexity.

Subjective ratings confirmed the objective results. Furthermore, using ratings that weighted the importance of the bipolar rating scale produced a metric that could distinguish workload within one of the three classes of flight tasks. Therefore, this improved subjective scale was more sensitive than the objective measure which could not discriminate within a class. Due to the short duration of each flight task, it is unlikely that the superiority of the weighted rating scale can be attributed to its measuring peak, rather than average, workload as suggested by Kantowitz et al (1983). Instead, weighted ratings may just be more sensitive measures. The use of such rating data is acceptable when confirmed by objective results.

The next step is to repeat this experiment using flight scenarios that combine more complex flight demands. Thus, instead of one of the present Base tasks, e.g., fly at constant speed, we would substitute a tracking task, e.g. VOR tracking. Then the corresponding Paired level would require VOR tracking while maintaining constant speed. Finally, an analog to the present Complex level would require VOR tracking, constant speed and controlled descent. We would anticipate results similar to the present with greater objective and subjective workload associated with increasing task compexity.

References

Johnson, D.F., Haywood, R.C., & Olson, W.M., 1982, Yoked Design and Secondary Task in Adaptive Training, Proceedings of the 26th Annual Meeting of the Human Factors Society. Los Angeles, CA, 21-24.

Kantowitz, B.H., Hart, S.G., Bortolussi, M.R., 1983, Measuring Pilot Workload in a Moving-Base Simulator: I. Asynchronous Secondary Choice-Reaction Task, Proceedings of the 27th Annual Meeting of the Human Factors Society at Norfolk, Virginia, 319-322.

Kantowitz, B.H., & Knight, J.L., 1976, Testing Tapping Timesharing II. Auditory Secondary Task., Acta Psychologica, 40, 343-362.

Wierwille, W.W. & Conner, S.A., 1983, Evaluation of 20 Workload Measures using a Pychomotor Task in a Moving-Base Simulator., Human Factors, 25, 1-16.

Williges, R.C. & Wierwille, W.W., 1979, Behavioral Measures of Aircrew Mental Workload, Human Factors, 21, 549-574.

Acknowledgment

This research was supported by Cooperative Agreement NCC 2-228 from the National Aeronautics and Space Administration, Ames Research Center. S.G. Hart was the Technical Monitor.

N85
14567

UNCLAS

MULTI-CREW MODEL ANALYTIC ASSESSMENT OF LANDING PERFORMANCE AND DECISION-MAKING DEMANDS

Paul Milgram, Rob van der Wijngaart, Henk Veerbeek
National Aerospace Laboratory NLR
Anthony Fokkerweg 2, 1059 CM Amsterdam
Okko F. Bleeker
Fokker B.V., POB 7600, 1117 ZJ Schiphol

ABSTRACT

Some of the relative merits of the PROCRU approach to modelling multi-crew flight deck activity during approach to landing are discussed. On the basis of two realistic flight scenarios, the ability of the model to simulate different vectored approaches is demonstrated. A second exemplary analysis of a nominal and an accelerated final approach is performed, illustrating the potential of the expected net gain (ENGP) functions as a measure of decision-making load.

1. INTRODUCTION

The ability to analyze various aspects of crew activity during the carrying out of well defined flight scenarios is of great potential value, both as a developmental tool for the design of flight decks and as an aid for ultimately approving their airworthiness. Included among the various issues to be addressed in this respect are the quantity, quality and synchronization of (electronic) display information, minimum equipment lists, minimum crew complements, normal operating and emergency procedures, system safety and reliability assessments, failure analyses and crew workload evaluation. Because the certificating of modern automated flight decks is in itself also a very complex undertaking, demanding as insightful and methodical an approach as possible, similar objective analysis tools are necessary for this task as well. Ideally, in fact, development and certification efforts should proceed hand in hand throughout all of the various engineering evaluation, testing and design review stages. As cockpit technology becomes more integrated and the pilot increasingly assumes the role of system supervisor, the logical trend is to commence such joint development and certification efforts ever earlier in the project planning and design stages.

Conventionally, a number of (overlapping) approaches have been adopted for predicting and evaluating crew performance and "workload" during early stages of design and for "measuring" it during the later stages. These include applying both objective performance measures and subjective expert opinion ratings to the evaluating of limited scale (laboratory) experiments, (static or dynamic) mockup assessments, full-scale (fixed- or moving-base) simulations and, ultimately, prototype flight tests. In the less advanced design stages, that is, before mockups, simulations or test flights have become (economically) feasible, task analytical methods may be used in order to indicate obvious design shortcomings and potential operational problems.

Analytical methods can be categorized in many ways, one of which is to distinguish between the use of closed-form solutions and the use of Monte Carlo numerical simulation methods. A large portion of the modelling efforts in the past decades, especially in the realm of control theoretic models, has

PRECEDING PAGE BLANK NOT FILMED

concentrated on the class of closed-form solutions, which renders predictions of stochastic ensemble-average performance measures. These models have a clear economical advantage over those which generate and then accumulate individual simulation runs in order ultimately to derive similar ensemble averaged measures. Closed-form modelling efforts have in general been limited, however, to the modelling of well defined, constrained and continuous task performance. In order to model more complex multi-task, multi-crew mixed discrete-continuous task performance, it is clear that important considerations with respect to system linearity, stationarity and ergodicity must carefully be taken into account before closed-form ensemble average prediction models may be employed. Although the importance of 'average' measures has certainly not decreased, it has become increasingly useful to employ numerical simulations of individual time lines as a human performance evaluation technique.

Another means of categorizing analytical methods is via bottom-up vs. top-down modelling. The majority of operational time-line analysis methods to date fall under the bottom-up classification and have been implemented by means of a variety of so-called "network models". In bottom-up modelling, human performance is synthesized from a sequence of fundamental activities, such as memory recalls, control actions, etc. (For discussion of the properties and relative merits of network models, the reader is referred to (1).) Two significant weaknesses of such methods are the difficulties frequently encountered during model checkout and validation, due to the intrinsic interconnectedness of the various task sub-models, and their stong dependence upon and sensitivity to the quality of data which must be gathered for setting model parameters. With respect to the latter, especially in modern integrated cockpits where humans perform primarily as system supervisors, the gathering of reliable behavioural data which relate to what the pilots are actually 'doing' is clearly very difficult.

In contrast to bottom-up methods, top-down models commence with a description of the task environment (system) that includes goals and sub-goals and then attempt to characterize the human component in modular fashion at the task or function level (1). More often than not such models are implemented normatively, that is, in a manner whereby the behaviour of the operator (and thus his/her performance) is prescribed according to a set of normative rules which are explicitly set up in order to optimize system performance with respect to its goals and sub-goals. "Validation" of a particular (sub)model, therefore, involves ascertaining whether or not the predicted performance measures, determined by a particular set of (sub)task objectives, match the corresponding operational performance measures. This is in contrast to the validating of strictly bottom-up models, which require that the performance of each model subcomponent agree with that of its structural analogue. Furthermore, because normative (sub)models need not necessarily be functionally anthropomorphous (although this is highly desireable), the reliability of their predictions is much less dependent upon empirical data.

The PROCRU model of flight crew procedure-oriented behaviour during commercial ILS approach-to-landing scenarios is a top-down model which combines elements of both normative and non-normative modelling, in order to generate numerically simulated time-lines of both continuous and discrete activity and of various covert and overt behavioural measures. The normative elements of PROCRU comprise well established optimal control, estimation and decision theoretic 'submodels' of human performance, which have been implemented in order to simulate the continuous regulating, monitoring, information processing and decision making behaviour of two interacting crew members: a pilot flying (PF) and a pilot not flying (PNF). Because most of the discrete tasks associated with an ILS approach to landing, which do not easily lend themselves to continuous modelling, are well defined routines which must be executed under well defined, event-related conditions, it has been possible to simulate the actual carrying out of such discrete procedural activity by means of purely functional modelling and to integrate this discrete behaviour with the associated continuous sub-task behaviour.

For a detailed description of the concepts underlying the PROCRU model and of the details of its implementation, the reader is referred to (2). Other more global descriptions of the model may be found in (1) and (3). The objective of this paper is to illustrate some of the top-down normative and non-normative aspects of the PROCRU model by providing a demonstration of some of the means by which PROCRU may be employed operationally in order to evaluate both overt crew activities during approach to landing and the normative decision making demands placed on the pilots while carrying out these activities.

2. EVALUATION OF LANDING PERFORMANCE

In order to illustrate PROCRU's ability to simulate outer loop landing performance, a short exemplary analysis has been carried out on the basis of a realistic landing scenario. The aircraft simulated is a Fokker F-28, manually operated by a 2-person crew. Aircraft and pilot parameters have been specified according to reference (2). The mission simulated is a standard ILS approach to landing on Runway 19R of Schiphol International Airport in The Netherlands, coming from the south. The approach is radar vectored and is performed under IFR, CAT II weather conditions, with a 150 foot cloud cover. No other traffic has been simulated and aircraft conditions are normal. In order to illustrate the effect of different levels of intensity of procedural activity, two versions of this basic mission scenario have been simulated: a long-turn approach and a short-turn approach. The ground tracks of these two approaches have been superimposed upon the map given in Fig. 1.

Both the mission scenarios and the outer loop PROCRU simulation results are summarized in Fig. 2 and Table 1. The latter appear in the form of major flight "milestones". Once the airport, aircraft, crew and procedural parameters have been set, the course of each scenario is determined by the air traffic control (ATC) vectors. Except for two changes the vectors used to generate the short-turn approach in Fig. 2B are the same as those for the long-turn approach in Fig. 2A.

Comparing Table 1A (long-turn) and 1B (short-turn) we see that the two scenarios correspond up to and including milestone no. 4. In the short-turn approach the crew is instructed at time t=290 s to decelerate in addition to the turn to 360°. This deceleration, which is not encountered in the long-turn approach, causes a decrease in the aircraft's velocity in the x-direction. Because in the long-turn approach the aircraft is travelling faster when the "decelerate to 180 Kts" command is given, the "Flaps 11" request (milestone no. 9) which is triggered by v=180 Kts occurs 76 s later than in Table 1B. At milestone no. 10 the crew is instructed in the long-turn approach to turn to 145° and decelerate to 140 Kts, whereas the short-turn initial approach phase is speeded up by instructing deceleration to 160 Kts instead.

In the long-turn scenario there follows a period of straight and level flight before localizer (LOC) intercept (milestone no. 12) at 2.5 dots. After the turn to localizer (190°) there follows another period of straight and level flight (2000 ft, 140 Kts) until the glideslope (GS) is intercepted at

2.9 dots, milestone no. 16. In the short-turn scenario, on the other hand, immediately after the turn to 145° and deceleration to 160 Kts the localizer is intercepted, also at 2.5 dots. Similarly, immediately upon completion of the turn to localizer the glideslope is intercepted, at 1.2 dots deviation. This leaves comparatively little time for making the prescribed gear request (at -1.0 dots) and then commencing the flare to GS (-2.5°). Immediately prior to crossing the GS a "Flaps 25" request is initiated and deceleration to 137 Kts is commenced. In spite of these rapidly successive manoeuvres the aircraft in the short-turn scenario establishes itself successfully on the LOC and GS and the final approach which follows is essentially identical to that of the long-turn scenario.

This example illustrates the capability within PROCRU to direct a simulated aircraft by means of (preprogrammed) ATC vectors at different rates along different nominal trajectories, characterized by a set of altitude, velocity, heading angle and path angle profiles. The respective profiles for the short- and long-turn scenarios simulated here are illustrated in Fig. 3. The method by which these trajectories are flown is such that at any moment in time the aircraft is flown along one of five nominal trajectory segments, corresponding to five standard, well-defined manoeuvres: Straight & Level flight, Deceleration, Turn, Flare and Descent. The pilot flying (PF) computes and (manually) implements the deterministic nominal trim control vector necessary for carrying out each manoeuvre. The PF is responsible also for regulating out random disturbances; this task is modelled by means of the Optimal Control Model. Systematic changes in aircraft dynamics corresponding to aerodynamic changes along the approach trajectory are also included in the model.

From the above it is clear that one should not underestimate the computational complexities involved in obtaining the nominal trajectories illustrated in Fig.'s 1-3. More importantly, it is imperative to realize that these trajectories are "flown" not by a complex but nevertheless deterministic trim-computing algorithm, but by a "stochastic" pilot, who has perceptual limitations, who can only monitor one display cluster at a time, who must perform a number of other important prescribed flight procedures and who must decide whether or not the carrying out of the required nominal manoeuvre is at all the most important action to be undertaken at a particular moment. This latter aspect of PROCRU is dealt with in the following section.

3. EVALUATION OF DECISION-MAKING DEMANDS

Whereas section 2 illustrates those aspects of the model which determine the outer loop flying of the airplane, in the present section we look at the underlying goal-oriented decision-making processes which govern that flying behaviour as well as other procedural activities. In order to accomplish this two more landing scenarios have been defined, focussing on the final approach sections only. Since the two final approaches in the first example are almost identical (Fig.'s 1-3) one of these (the short-turn scenario) was chosen as the nominal scenario for the present example and a new scenario for an 'accelerated' final approach was defined.

The ground tracks for these two scenarios are shown in Fig. 4. The accelerated approach was created by initiating the simulation immediately prior to the localizer intercept, with a velocity of 140 Kts, heading 160° and altitude 1500 ft. As was shown in section 2, the aircraft in the nominal case has sufficient time at 2000 ft to complete the turn to localizer (190°) and functions. These functions assign a measure of "urgency" to each procedure at any moment in time, on the basis of which priority for execution of that procedure is decided. This is done rather elegantly, in terms of both the static and dynamic gain of the procedure.

The static ENGP value is a constant which is chosen to reflect the baseline value associated with the procedure. It serves the dual purpose both of establishing a default activity for each crew member, as discussed earlier, and of deciding between two or more procedures which may happen to be equally relevant at any particular time. The dynamic part of the ENGP function is designed to represent the "situational relevance" of each procedure. The "situational" aspect is determined by a function, presently a unit step function, which enables its corresponding procedure following the occurrence of an appropriate event. The "relevance" aspect of the dynamic ENGP function is either a function of the pilot's stochastic internal estimate of a relevant state variable or an explicit function of external events. The former class of functions is modelled in PROCRU by means of two types of "appropriateness" functions, which become non-zero contingent upon exceedence of either a preset critical level or a tolerance window. The latter are implemented by means of a "timeliness" function, which weighs the gain for executing a procedure as a function of the elapsed time following a predicating event.

The ENGP functions clearly fulfill an important role in the normative PROCRU model. That is to say, the choosing of procedures for execution on the basis of maximizing expected net gain corresponds to the goal-directed behaviour associated with top-down modelling discussed earlier. In situations wherein the individual gains for executing two or more procedures conflict, it is possible to postulate that the total decision-making load imposed upon the pilot(s) increases accordingly.

In Fig.'s 6B-9B this postulate is illustrated in the plots of the ENGP functions corresponding to the procedures in Fig. 6A-9A. In Fig. 6B it is clear that the PF undergoes little decision-making load. In Fig. 7B, however, we see that the ENGP function for default procedure no. 26, Flying the Airplane, has risen above its minimum constant value (0.3), in contrast to procedure no. 26 in Fig. 6B. The behaviour of this function is the net result of the glideslope overshoot discussed earlier (see Fig. 5). Because the probability of the state vector components' being within tolerance is decreased in this case, ENGP appropriateness function no. 26 is correspondingly larger. This gives rise to a number of minor conflicts with other ENGP functions, namely the retrim porcedure (no. 4) and message decoding (no. 1), which have risen to levels higher than those attained in Fig. 6B.

Similar conclusions to the above may be derived from the ENGP plots for the PNF in Fig. 8B and 9B. In Fig. 9B the ENGP appropriateness function no. 22 (monitoring aircraft status) also achieves levels much higher than in Fig. 8B, due to the same consequences of the glideslope overshoot. The conflict situation which we observe here is that, because ENGP function no. 22 is relatively high, ENGP function no. 21 (approach stability monitoring) also increases accordingly, since procedure no. 21 cannot be executed until ENGP(21) is greater than ENGP(22). Fig.'s 8B and 9B are furthermore especially interesting when recalling the procedure time-lines in Fig.'s 8A and 9A. There it was noted earlier that it is difficult to discern between the oscillatory behaviours in each figure. On the basis of the ENGP functions, however, much more insight is now available; although the (covert) performance of the two supervisory PNF's is similar in the two scenarios, it is now possible to speculate that their behaviours are indeed different.

then commence the flare to glideslope (-2.5°). In the accelerated approach, however, the turn to localizer is begun at 1500 ft with localizer deviation 2.2 dots. Immediately after the turn is commenced the glideslope becomes alive at 1.0 dots. During the turn, and in quick succession, the GS alive callout is made and then processed, the final approach checklist is requested and then started, the gears are requested and lowered, Flaps 25° are requested and the outer marker (OM) becomes active! When the turn is completed the flare to glideslope is commenced with the aircraft already 0.3 dots above the glideslope. The resulting overshoot reaches a maximum of 1.0 dots, but is slowly corrected over a period of approximately 90 s. As shown in Fig. 5 the two sets of profiles eventually converge just prior to touchdown and safe landings are achieved in both cases.

In order to discuss more thoroughly the specific actions which have been carried out by the pilots in enacting these two landing scenarios, Fig.'s 6-9 are presented. In each of these figures the upper plots (Fig. 6A-9A) show the "procedure time lines" generated by PROCRU. Keeping in mind that in the model each pilot is occupied at all times with one and only one procedure, these plots consist of a series of binary 'pulses', whereby OFF (low) corresponds to inactive and ON (high) corresponds to procedure active. There are 27 of these procedure curves for the PF and 28 for the PNF. The procedure numbers corresponding to each procedure plot are shown along the vertical axis. In Table 2A and 2B lists of the actions and enabling conditions corresponding to each procedure are given for the PF and PNF respectively. (As evident in Table 2A and 2B, not all procedure numbers are implemented for the present scenarios; they are nevertheless all plotted in Fig.'s 6-9.)

Of special interest in these figures are the so-called "default" procedures, comprising those activities which are carried out by either crew member during periods in which no other activities are demanded. For the PF (Fig. 6&7) this is procedure 26: Flying the Airplane, and for the PNF (Fig. 8&9) this is procedure 22: Monitoring Aircraft Status. As evident in the figures, in contrast to the rest of the procedure time lines which are usually inactive and only on occasion active, the two default procedures exhibit the reverse pattern.

The types of conclusions which may provisionally be drawn from these time lines are relatively straightforward. By noting how densely spaced and periodically the (non-default) procedures occur, it is possible to derive (numerical) measures for quantifying "intensity of activity". Comparing Fig. 6A and 7A, for example, we see, as expected, that the "activity" of the PF appears to be slightly more intense during the accelerated final approach. In Fig. 8A and 9A the most conspicuous pattern is the periodic alternating between the default procedure 22 (Monitoring Aircraft Status) and procedure 21 (Monitoring Approach Stability) after the Outer Marker (OM) has been passed. It is not immediately obvious from these plots whether the intensity of PNF activity is greater for the nominal than for for the accelerated approach.

Although the information contained in the procedure time lines is of great potential value this information is nevertheless limited, due to the fact that only one procedure is active in the model at any one time. It is not possible, therefore, to derive directly from these plots insight into the information processing considerations which have led to the decisions to carry out particular procedures at particular moments in time. Consequently, it is also not easy to detect the occurrence of conflicts which may have arisen while these decisions were being made. Such information is available, however, from the so-called Expected Net Gain for Procedure execution (ENGP)

Finally, it is obvious that an efficient means of integrating the above ideas is to add up the ENGP functions over all of the procedures for each pilot and to produce a total summa ted ENGP function, which has many attractive aspects as a potential measure of total procedural-related decisionmaking demand. This notion is illustrated in Fig. 10, which shows the summated gain functions corresponding to Fig.'s 6-9. All of the points discussed earlier with respect to the postulated increase in decision-making load on both the PF and PNF during the accelerated approach are illustrated here very clearly.

4. CONCLUSIONS

The exemplary analyses and discussions in this paper demonstrate some practical operational applications of analytical modelling, in particular of the PROCRU procedure oriented crew model, for predicting problems and potential conflicts in flight crew decision-making during approach to landing. Because covert supervisory behaviour of human operators is difficult to measure, mathematical human performance modelling is expected to become increasingly important as modern flight decks become more integrated and the supervisors' tasks grow more complex. Although simulation time-lines have proven to be useful for investigating different aspects of crew behaviour, most traditional bottom-up methods possess a number of serious weaknesses. Top-down (normative) modelling, on the other hand, goes a long way towards alleviating these problems. The top-down PROCRU approach, as illustrated here, also retains a number of favourable aspects of functional bottom-up modelling, in order to generate time lines of both discrete and continuous flight task activity and of both overt and covert pilot behaviour.

REFERENCES

(1) RW Pew & S Baron (1983): Perspectives on Human Performance Modelling. Automatica, Vol. 19(6), 663-676.

(2) S Baron, R Muralidharan, R Lancraft, G Zacharias (1980): PROCRU: A Model for Analyzing Crew Procedures in Approach to Landing. NASA CR-152397.

(3) S Baron, G Zacharias, R Muralidharan, R Lancraft (1980): PROCRU: A Model for Analyzing Crew Procedures in Approach to Landing. 16th Annual Conf. on Manual Control (MIT), 488-520.

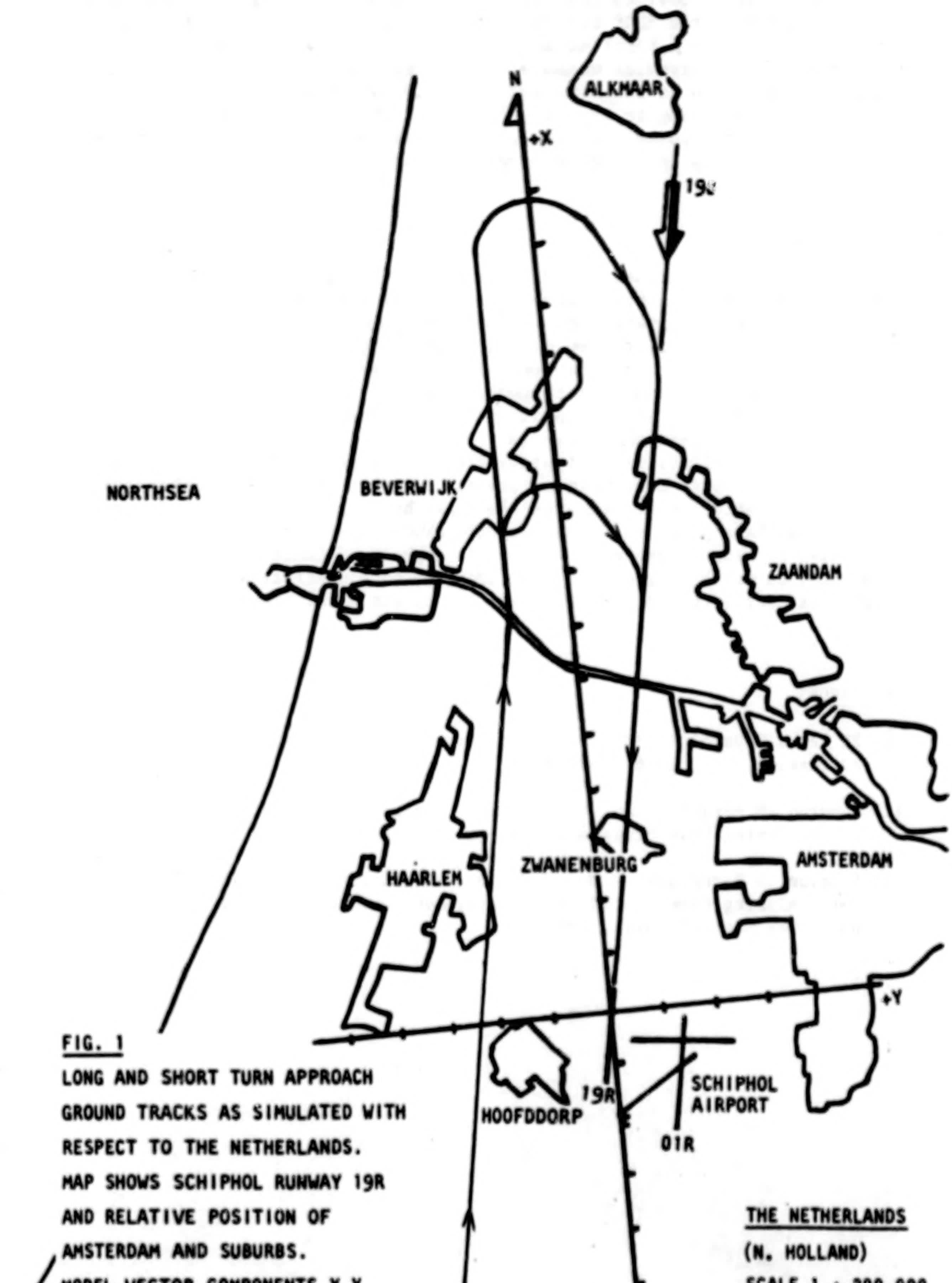

ALKMAAR
N
+X
19
NORTHSEA
BEVERWIJK
ZAANDAM
HAARLEM
ZWANENBURG
AMSTERDAM
+Y
19R
HOOFDDORP
SCHIPHOL
AIRPORT
01R
THE NETHERLANDS
(N. HOLLAND)
SCALE 1 : 200.000
= 6562 ft.

FIG. 1

LONG AND SHORT TURN APPROACH GROUND TRACKS AS SIMULATED WITH RESPECT TO THE NETHERLANDS. MAP SHOWS SCHIPHOL RUNWAY 19R AND RELATIVE POSITION OF AMSTERDAM AND SUBURBS. MODEL VECTOR COMPONENTS X,Y CORRESPOND TO N,E RESPECTIVELY.

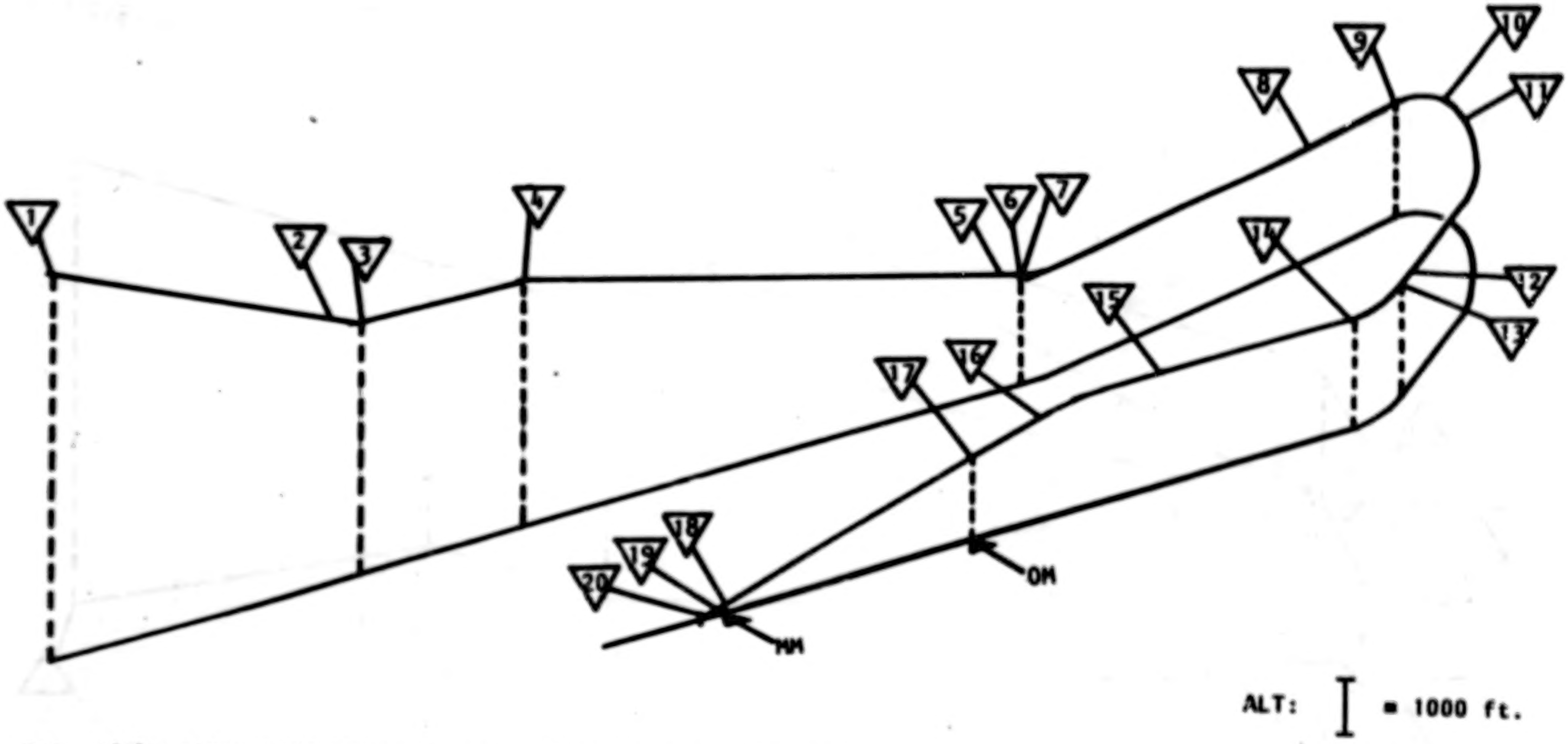

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
OM
MM
ALT: = 1000 ft.

FIG. 2(A) PROCRU LONG-TURN APPROACH TO SCHIPHOL RUNWAY 19R.
GROUNDTRACK WITH VERTICAL PROFILE;
NUMBERS REFER TO MAJOR MILESTONES IN TABLE 1(A)

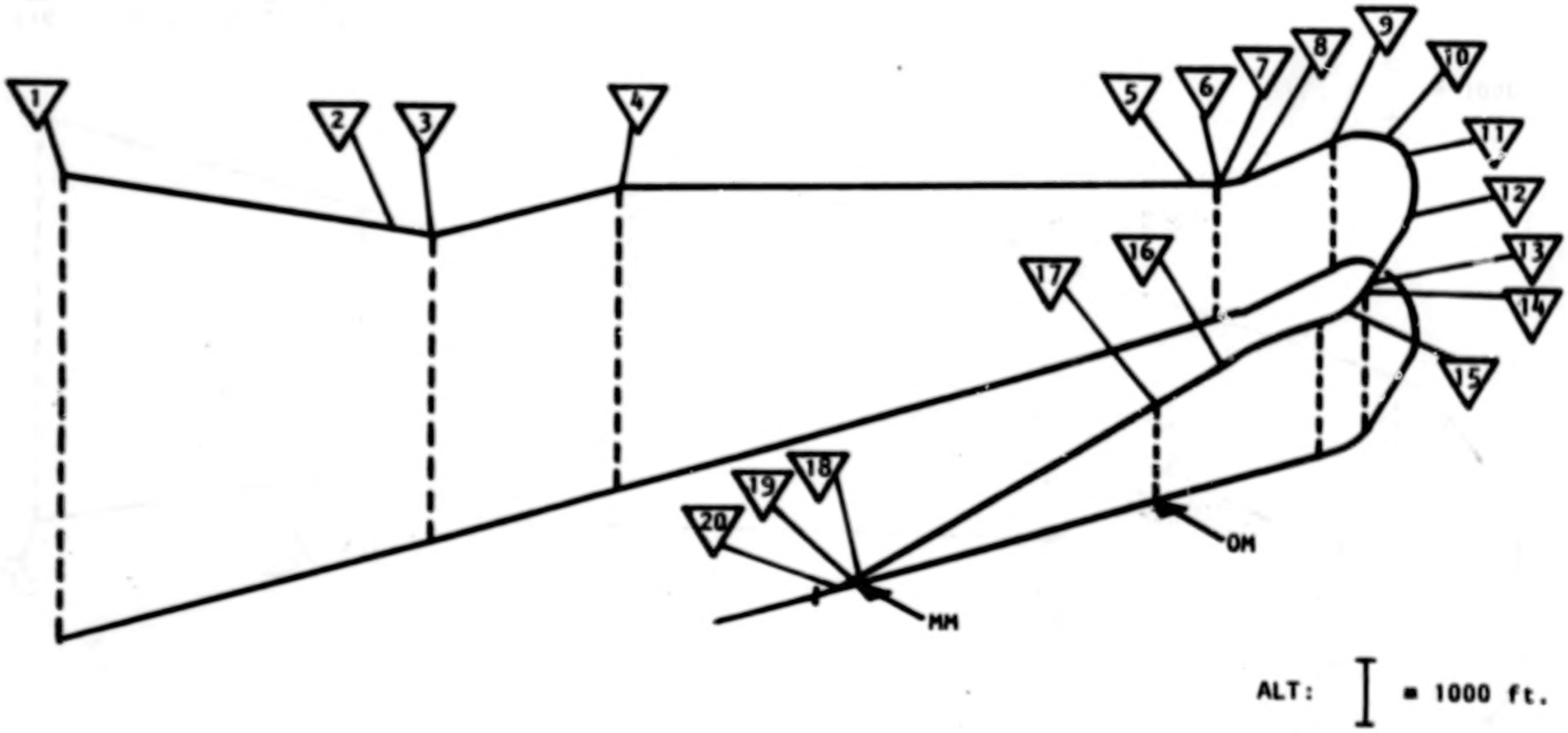

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
OM
MM
ALT: = 1000 ft.

FIG. 2(B) PROCRU SHORT-TURN APPROACH TO SCHIPHOL RUNWAY 19R.
GROUNDTRACK WITH VERTICAL PROFILE;
NUMBERS REFER TO MAJOR MILESTONES IN TABLE 1(B)

TABLE 1(A)

PROCRU Long-turn approach into Schiphol runway 19R; milestone summary.

no.	time (s)	Description
1	0	Altitude = 7000 ft., Velocity = 250 kts, Heading = 010° ATC vector: "Descend to 4500 ft."
2	83	ATC vector: "Decelerate to 210 kts; descend to 2000 ft."
3	85	Flares from flight path angle -4.1° to 0° at 4500 ft. to decelerate.
4	130	Flares from flight path angle 0° to -2.4° to descend to 2000 ft.
5	290	ATC vector: "Turn to 360°"
6	298	Flares from flight path angle -2.4° to 0° at 2000 ft.
7	300	Begins turn from 010° to 360°
8	400	ATC vector: "Decelerate to 180 kts, turn to 100°"
9	432	Begins turn from 360° to 100° (after the deceleration); PF requests initial flaps (11°)
10	443	ATC vector: "Turn to 145°; decelerate to 140 kts."
11	475	ATC: "Cleared for approach"
12	529	Intercepts localizer at 2.5 dots deviation
13	531	Begins turn to 190° to the localizer
14	573	Intercepts glideslope at -2.9 dots deviation
15	671	PF requests landing gear down
16	708	PF requests approach flaps (25°)
17	753	Outer Marker active
18	882	Middle Marker active
19	892	PNF calls: "Runway in sight"
20	903	Flares from flight path angle -2.5° to 0° in order to land

TABLE 1(B)

PROCRU Short-turn approach into Schiphol runway 19R; milestone summary.

No.	time (s)	Description
1	0	Altitude = 7000 ft., Velocity = 250 kts, Heading = 010° ATC vector: "Descend to 4500 ft."
2	83	ATC vector: "Decelerate to 210 kts; descend to 2000 ft."
3	85	Flares from flight path angle -4.1° to 0° at 4500 ft. to decelerate.
4	130	Flares from flight path angle 0° to -2.4° to descend to 2000 ft.
5	290	ATC vector: "Turn to 360°; decelerate to 190 kts"
6	298	Flares from flight path angle -2.4° to 0° at 2000 ft.
7	300	Begins turn from 010° to 360°
8	313	ATC vector: "Decelerate to 180 kts, turn to 100°"
9	356	PF requests initial flaps (11°)
10	356	ATC vector: "Turn to 145°; decelerate to 160 kts"
11	370	ATC: "Cleared for approach."
12	383	Initiates deceleration from 180 to 160 kts.
13	404	Intercepts localizer at 2.5 dots deviation
14	407	Intercepts glideslope at -1.2 dots deviation
15	417	PF requests landing gear down
16	449	PF requests approach flaps (25°) (glideslope dev. 0°)
17	493	Outer Marker active
18	621	Middle Marker active
19	632	PNF calls: "Runway in sight"
20	642	Flares from flight path angle -2.5° to 0° in order to land

ORIGINAL PAGE IS
OF POOR QUALITY

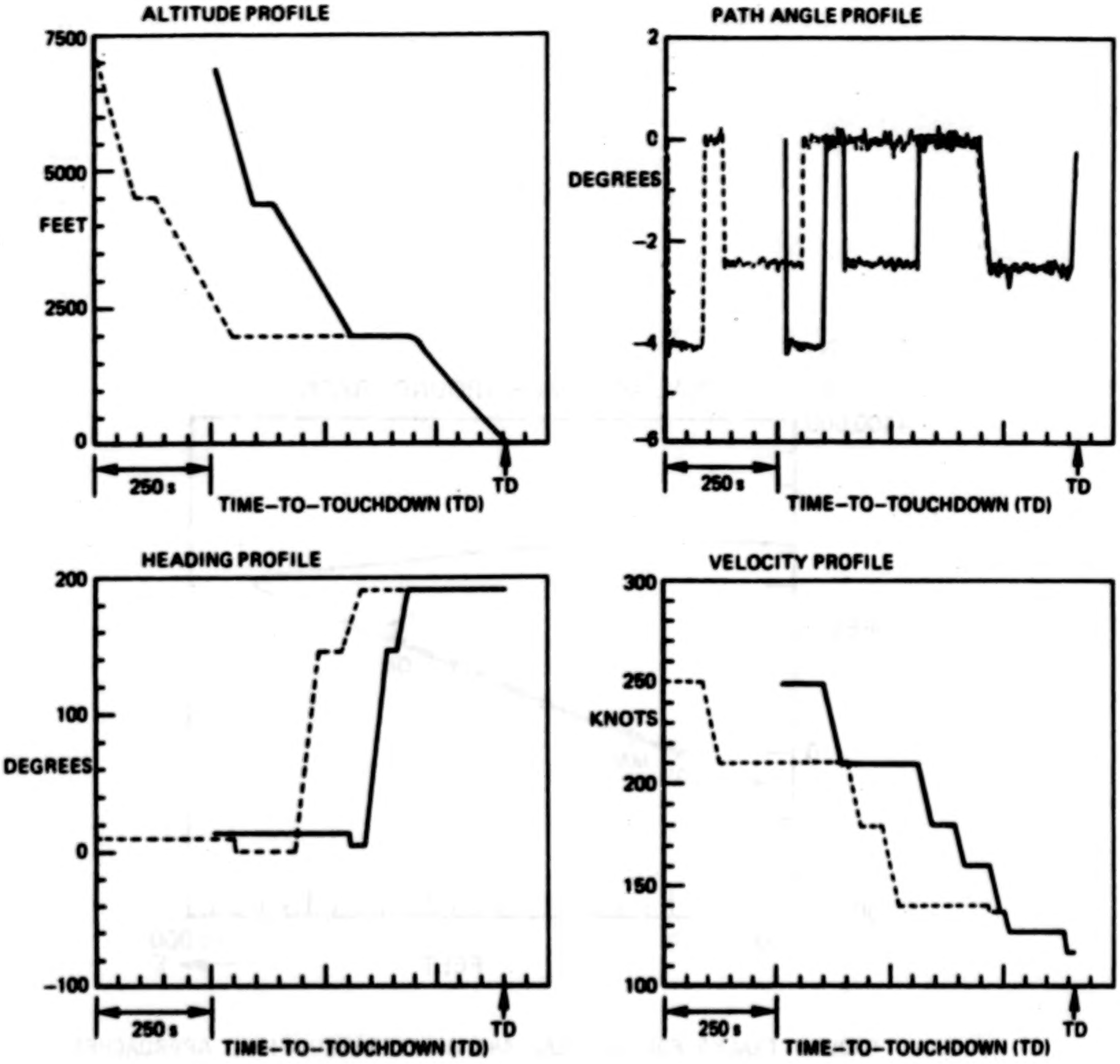


FIG. 3 FLIGHT PROFILES FOR SHORT- AND LONG-TURN SCENARIOS

- - - LONG-TURN
— SHORT-TURN

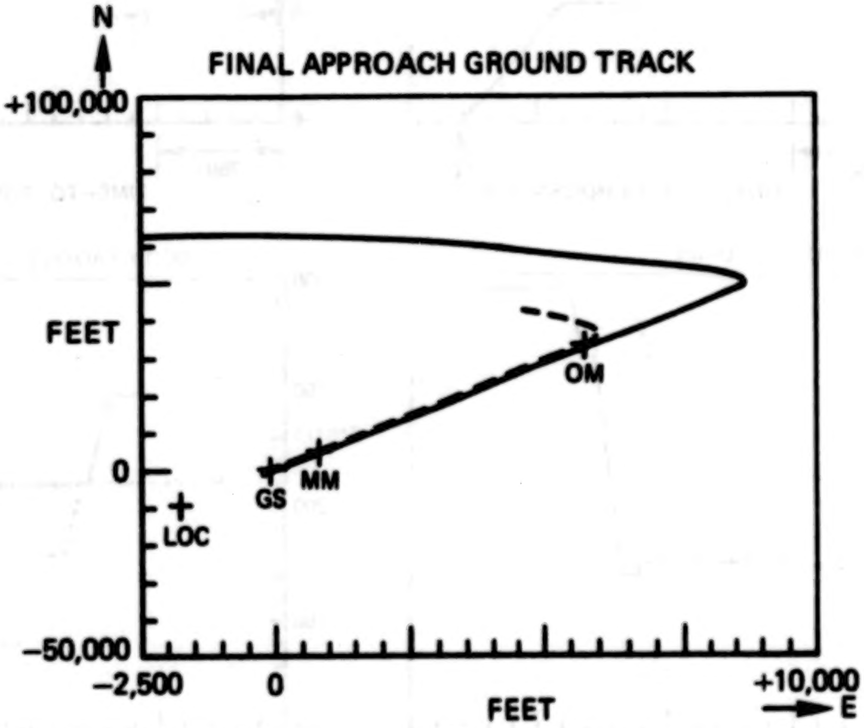


FIG. 4 GROUND TRACKS FOR NOMINAL AND ACCELERATED FINAL APPROACHES

—— NOMINAL (SHORT-TURN)

– – – ACCELERATED

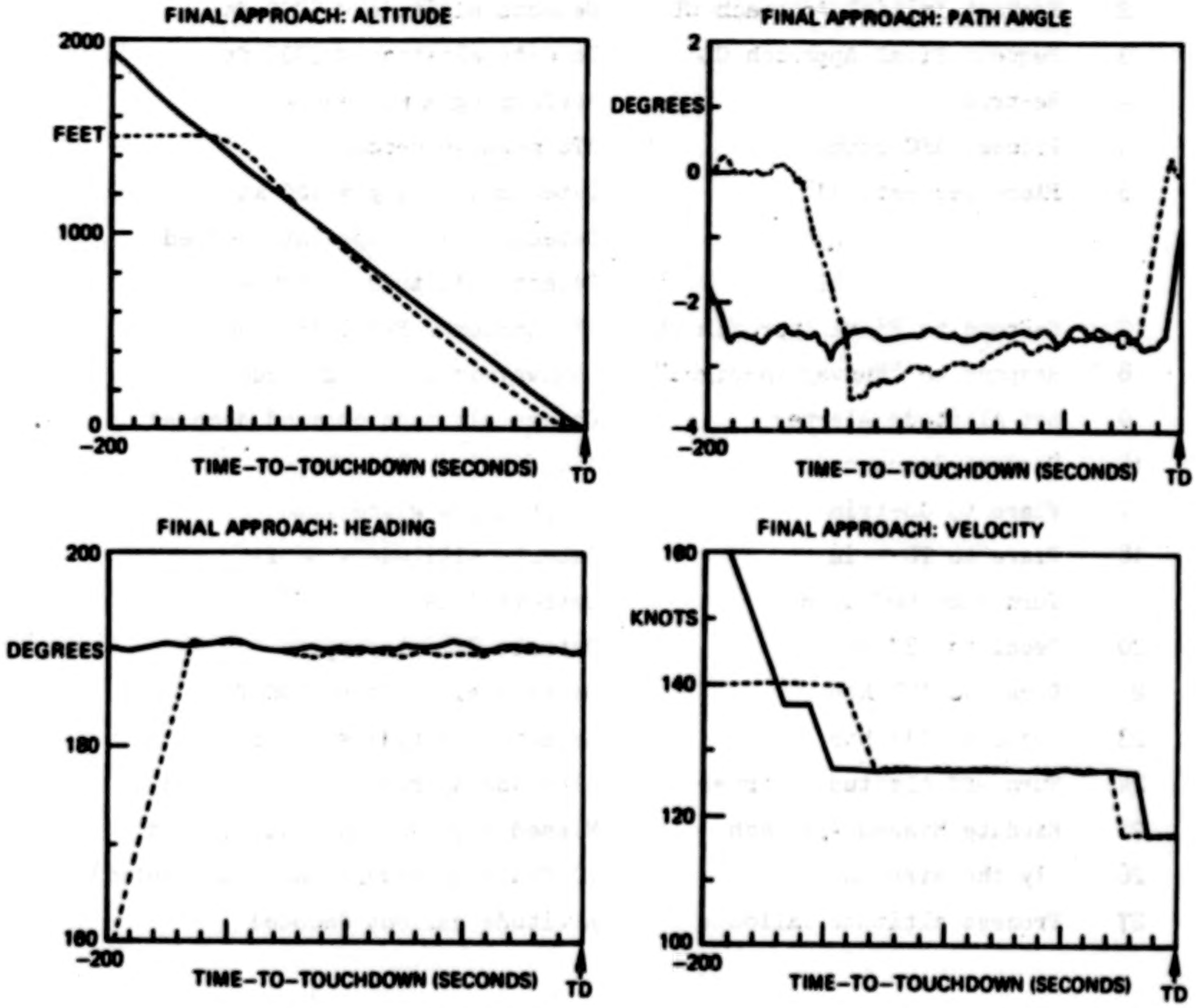


FIG. 5 FLIGHT PROFILES FOR NOMINAL AND ACCELERATED FINAL APPROACHES

—— NOMINAL (SHORT-TURN)

- - - ACCELERATED

TABLE 2A

Procedures for Pilot Flying (PF)

No.	Procedures	Enabling Condition
1	Decode message	Any audio activity
2	Request Initial Approach CL	Detects altitude < 5000 ft
3	Request Final Approach CL	Detects altitude < 2000 ft
4	Re-trim	Performing a manoeuvre
5	Process ATC command	ATC message decoded
6	Flaps request: 11°	Detects Velocity = 180 kt
	25°	Detects Glideslope intercepted
	42°	Detects Altitude = 1500 ft
7	Respond to Final Approach CL	PNF performs Final Approach CL
8	Respond to "Runway-in-sight"	"Runway-in-sight" decoded
9	Set Altitude Alerter	Change altitude command decoded
16	Request Gears-down	1 dot below glideslope
17	Flare to GS-trim	½ dot below glideslope
18	Flare to TD-trim	Detects altitude < 50 ft
19	Turn onto LOC-trim	Detects Localizer "ON"
20	Decel to 137 kts	Detects GS-intercepted
21	Decel to 127 kts	Detects altitude < 1500 ft
23	Decel to 117 kts	Detects altitude < 150 ft
24	Turn off Altitude Alerter	Altitude Alerter Alarm decoded
25	Execute Missed Approach	Missed Approach callout decoded
26	Fly the airplane	(Default procedure: always enabled)
27	Process altitude callouts	Altitude callout decoded

TABLE 2B

Procedures for Pilot Not Flying (PNF)

No.	Procedures	Enabling Condition
1	Decode message	Any audio activity
2	LOC Alarm callout	LOC Alarm decoded
3	GS Alarm callout	GS Alarm decoded
4	OM Alarm callout	OM Alarm decoded
5	MM Alarm callout	MM Alarm decoded
6	Begin monitoring external scene for Runway in Sight	Detects altitude = search height = cloud cover (150 ft) + 75 ft
8	Respond to Flaps request	Flaps request decoded
9	Respond to Gear request	Gear request decoded
10	Acknowledge ATC message	ATC message decoded
21	Monitoring approach stability	OM Alarm decoded
22	Monitoring aircraft status	(Default procedure: always enabled)
23	1000 ft altitude callout	Detects altitude = 1000 + 25 ft
24	500 ft altitude callout	Detects altitude = 500 + 25 ft
25	Approach minimum altitude callout	Detects altitude = 300 + 25 ft
26	Minimum altitude callout	Detects altitude = 200 + 25 ft
27	Perform Initial Approach CL	Initial Approach CL request decoded
28	Perform Final Approach CL	Final Approach CL request decoded

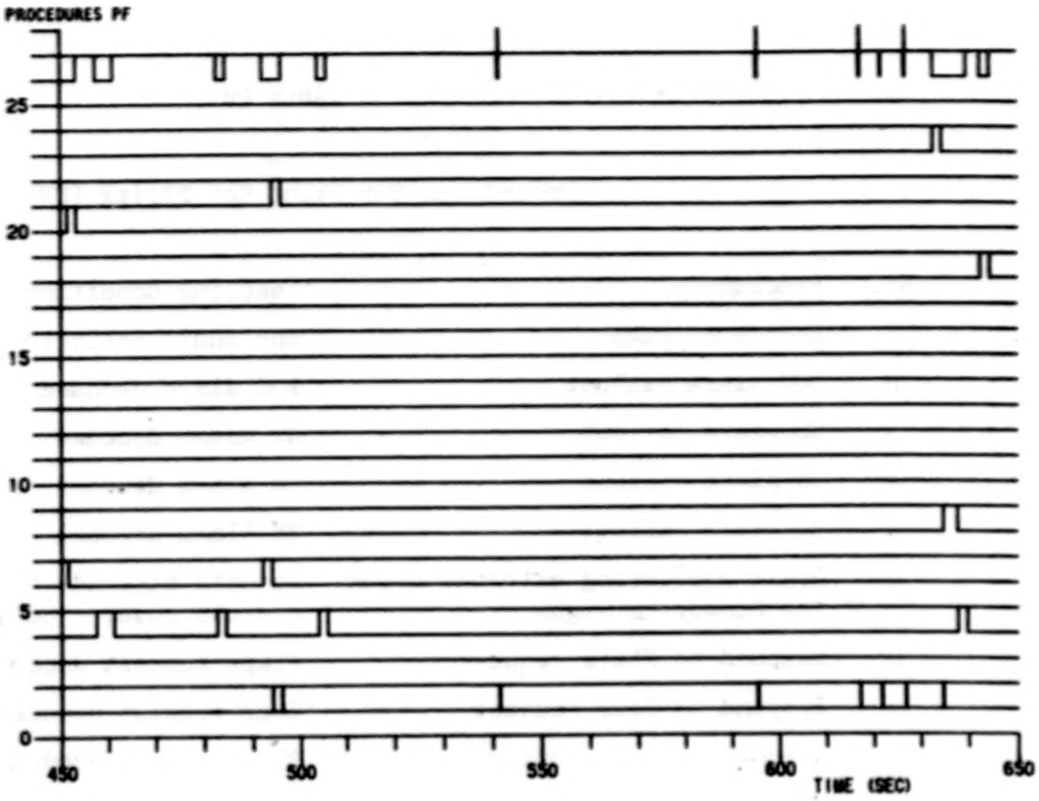


FIG. 6A PF PROCEDURE TIME-LINE FOR NOMINAL FINAL APPROACH (SEE TABLE 2A)

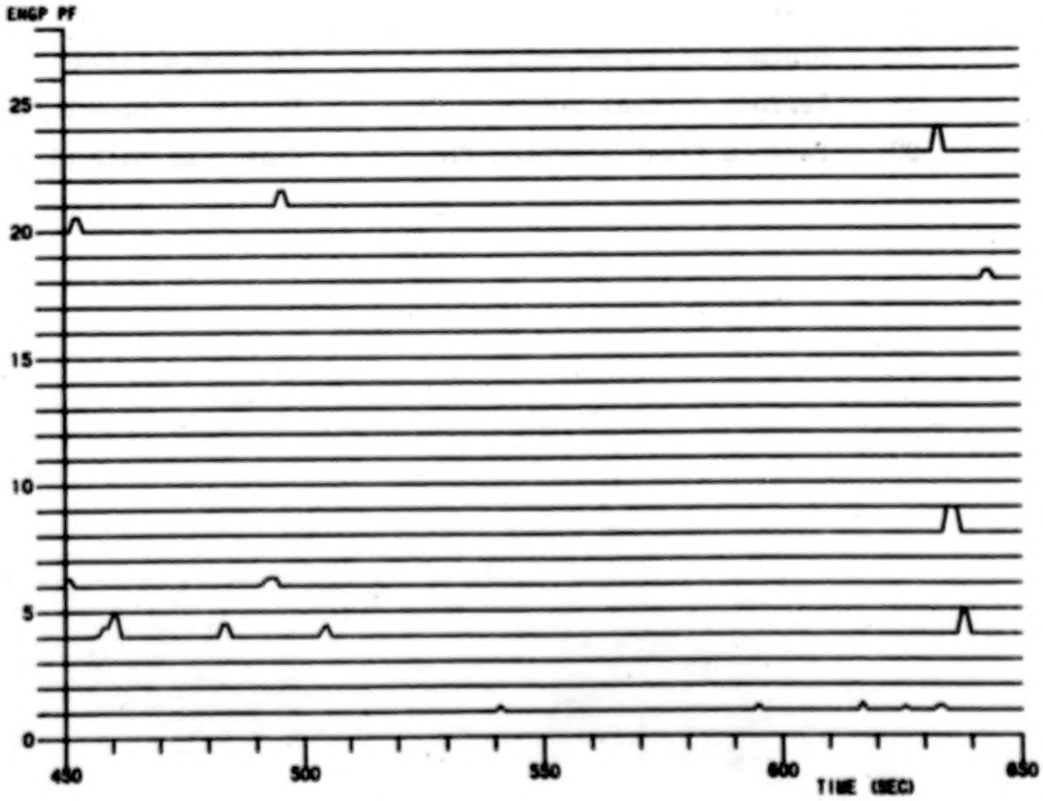


FIG. 6B PF GAIN (ENGP) FUNCTIONS FOR NOMINAL FINAL APPROACH (SEE TABLE 2A)

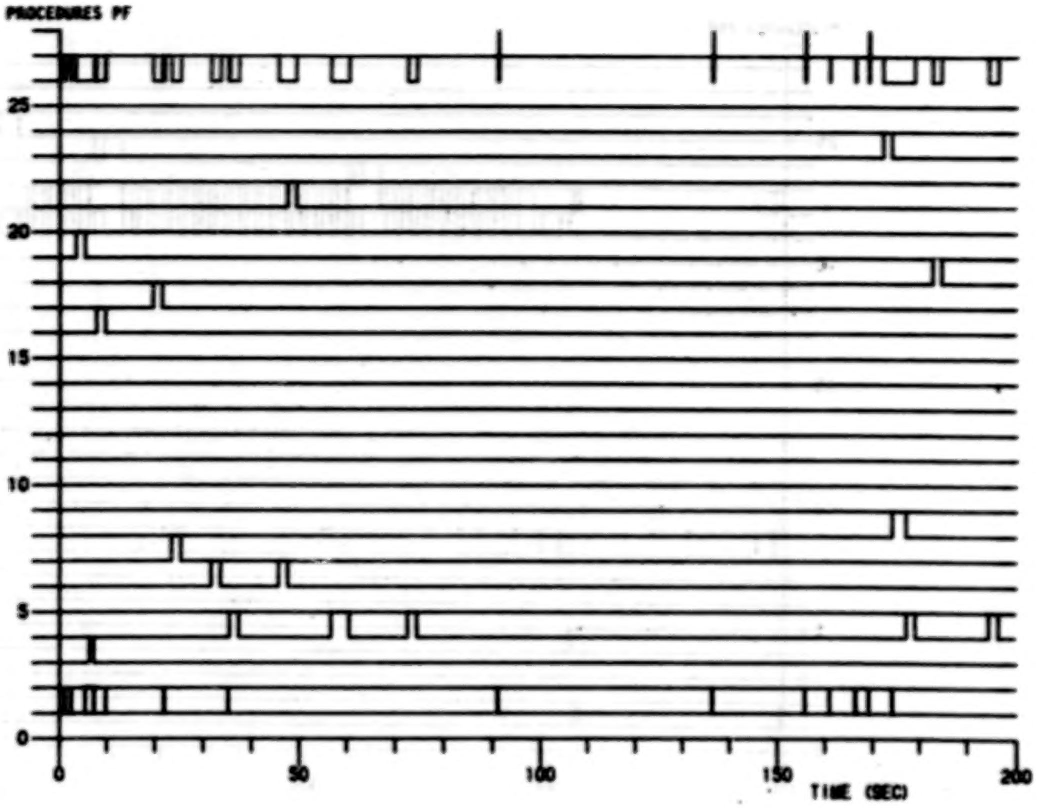


FIG. 7A PF PROCEDURE TIME-LINE FOR "ACCELERATED" FINAL APPROACH (SEE TABLE 2A)

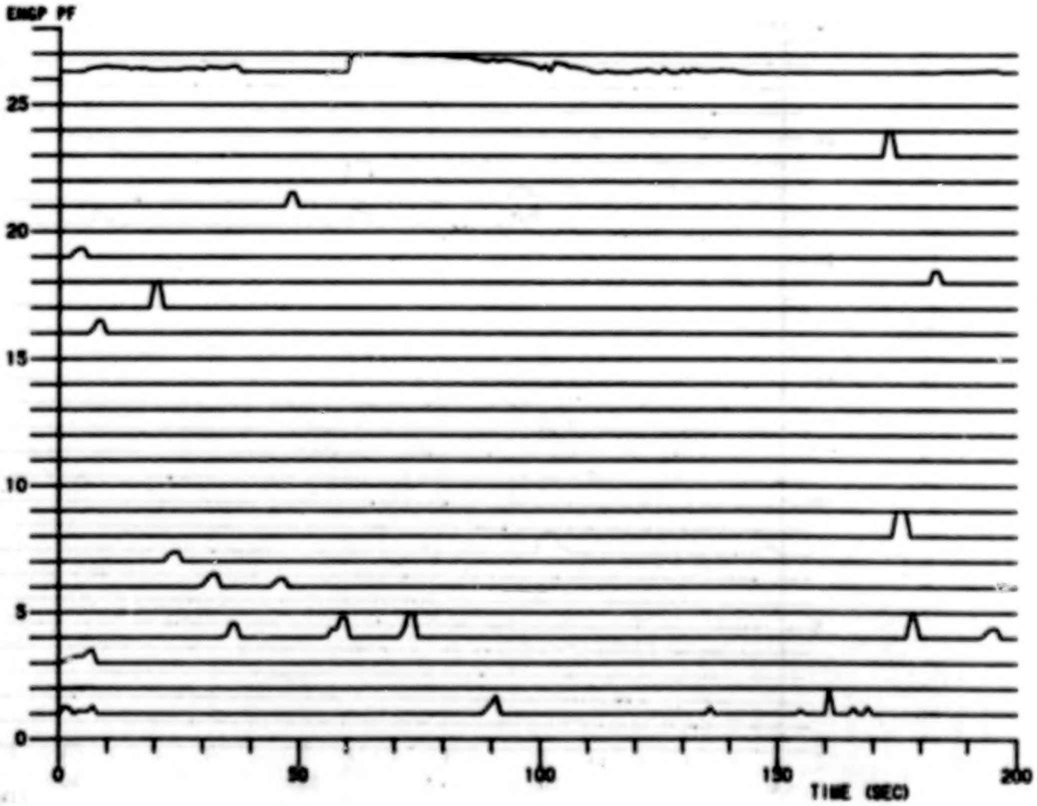


FIG. 7B PF GAIN (ENGP) FUNCTIONS FOR "ACCELERATED" FINAL APPROACH (SEE TABLE 2A)

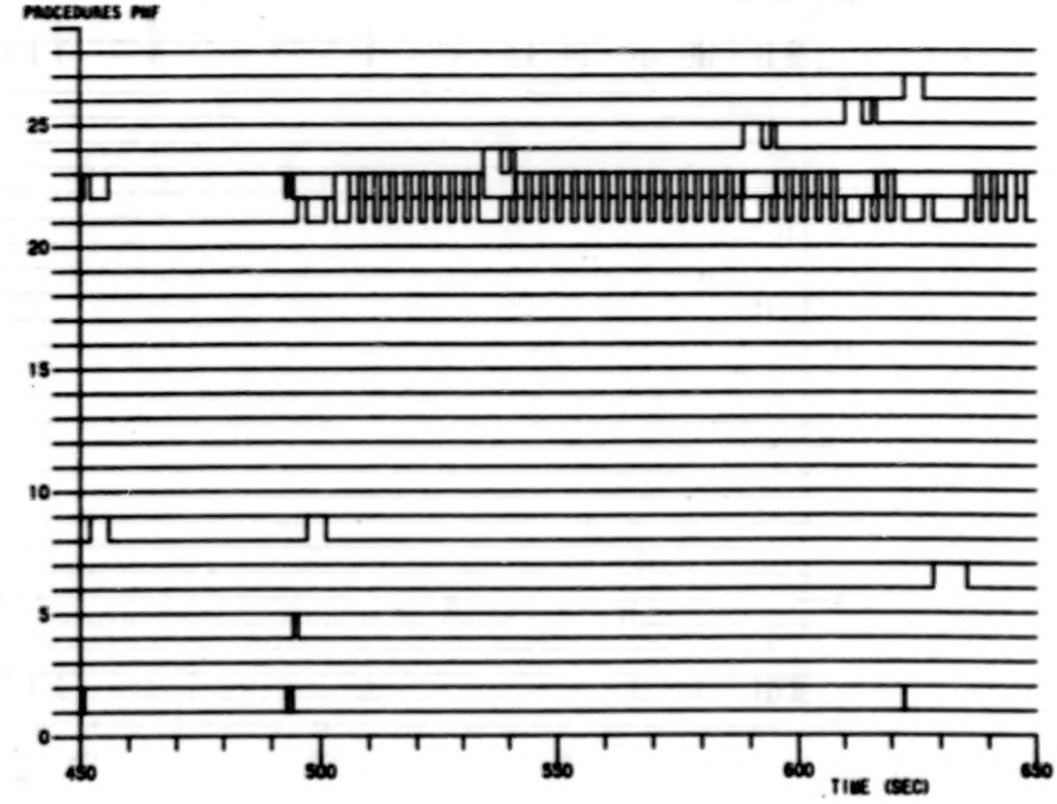


FIG. 8A PNF PROCEDURE TIME-LINE FOR NOMINAL FINAL APPROACH (SEE TABLE 2B)

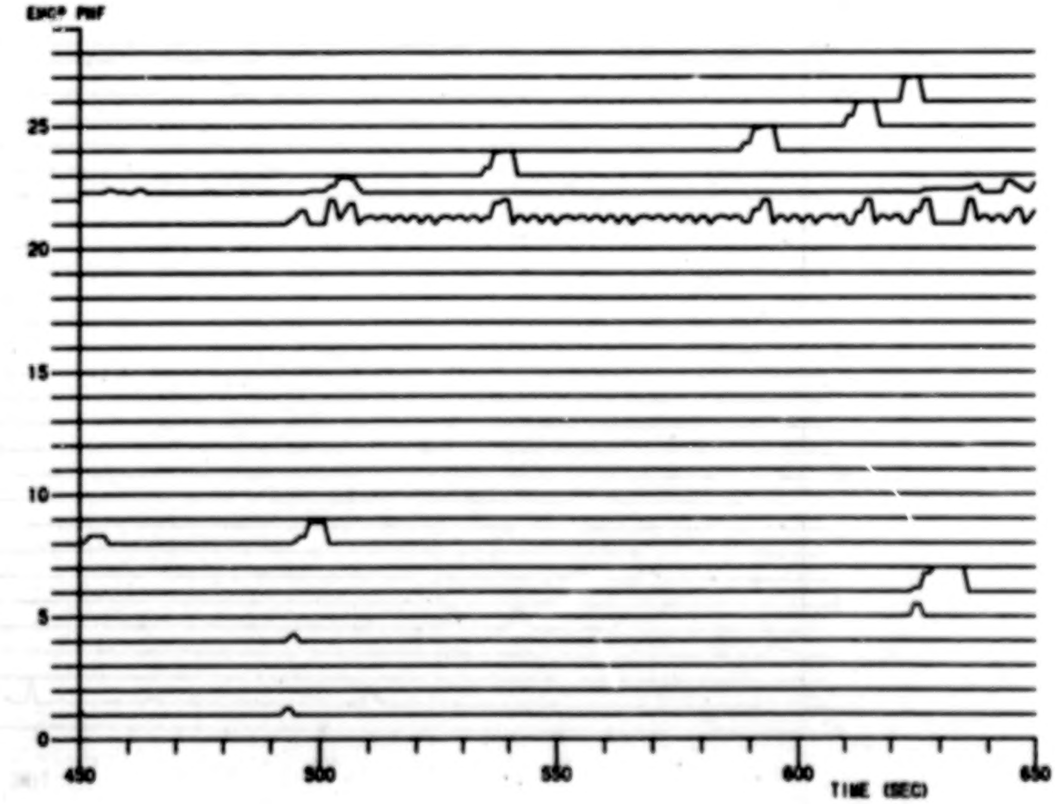


FIG. 8B PNF GAIN (ENGP) FUNCTIONS FOR NOMINAL FINAL APPROACH (SEE TABLE 2B)

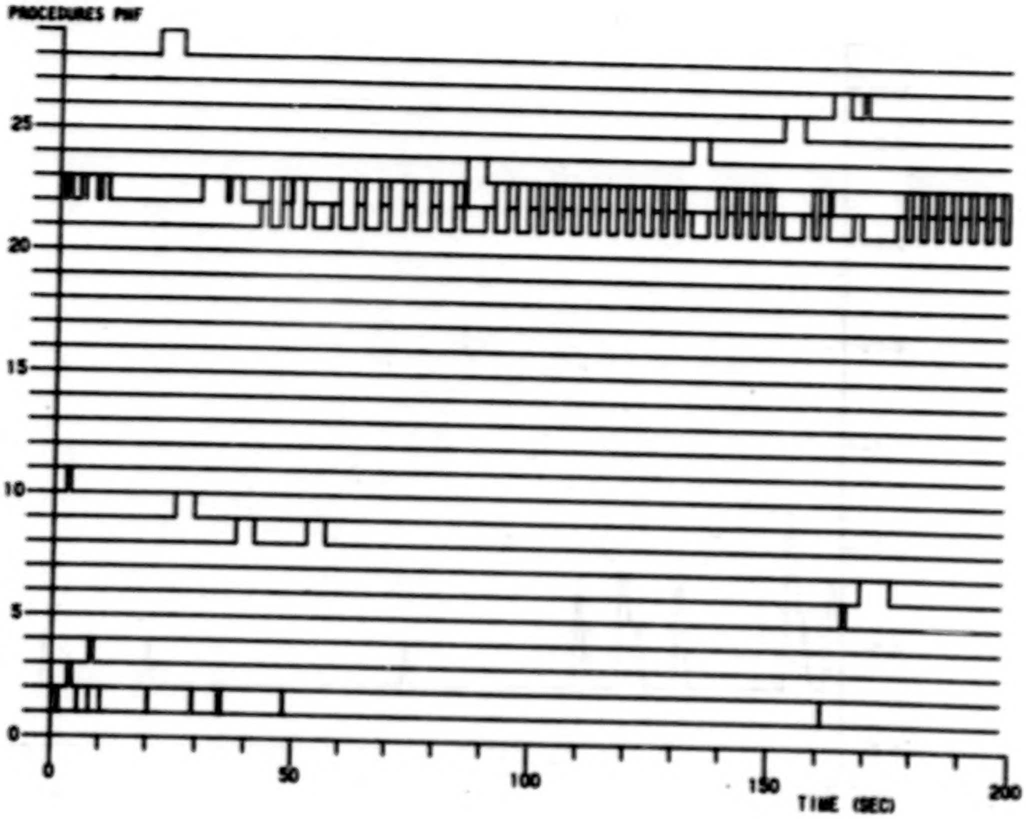


FIG. 9A PNF PROCEDURE TIME-LINE FOR "ACCELERATED" FINAL APPROACH (SEE TABLE 2B)

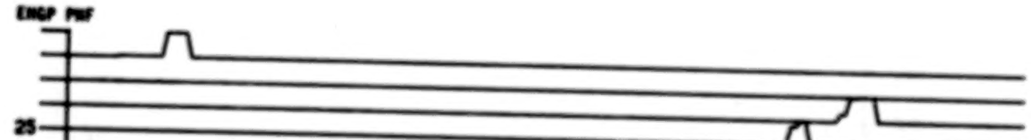


FIG. 9B PNF GAIN (ENGP) FUNCTIONS FOR "ACCELERATED" FINAL APPROACH (SEE TABLE 2B)

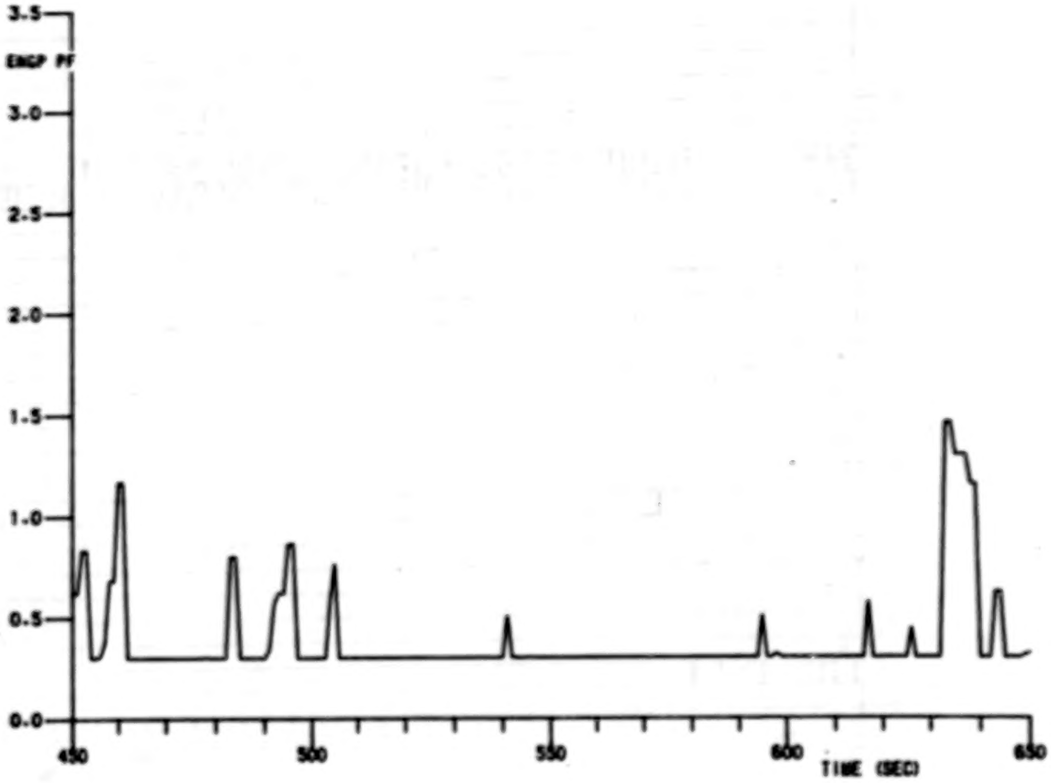


FIG. 10A SUMMATED PF GAIN (ENGP) FUNCTIONS FOR NOMINAL FINAL APPROACH

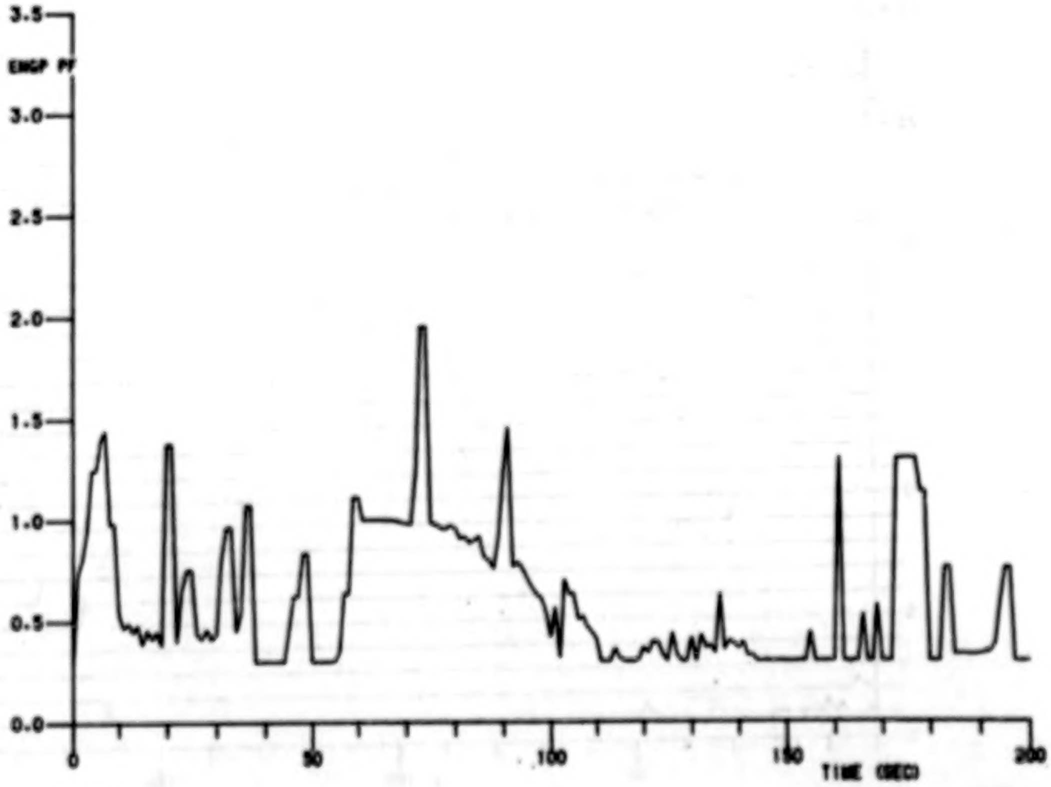


FIG. 10B SUMMATED PF GAIN (ENGP) FUNCTIONS FOR "ACCELERATED" FINAL APPROACH

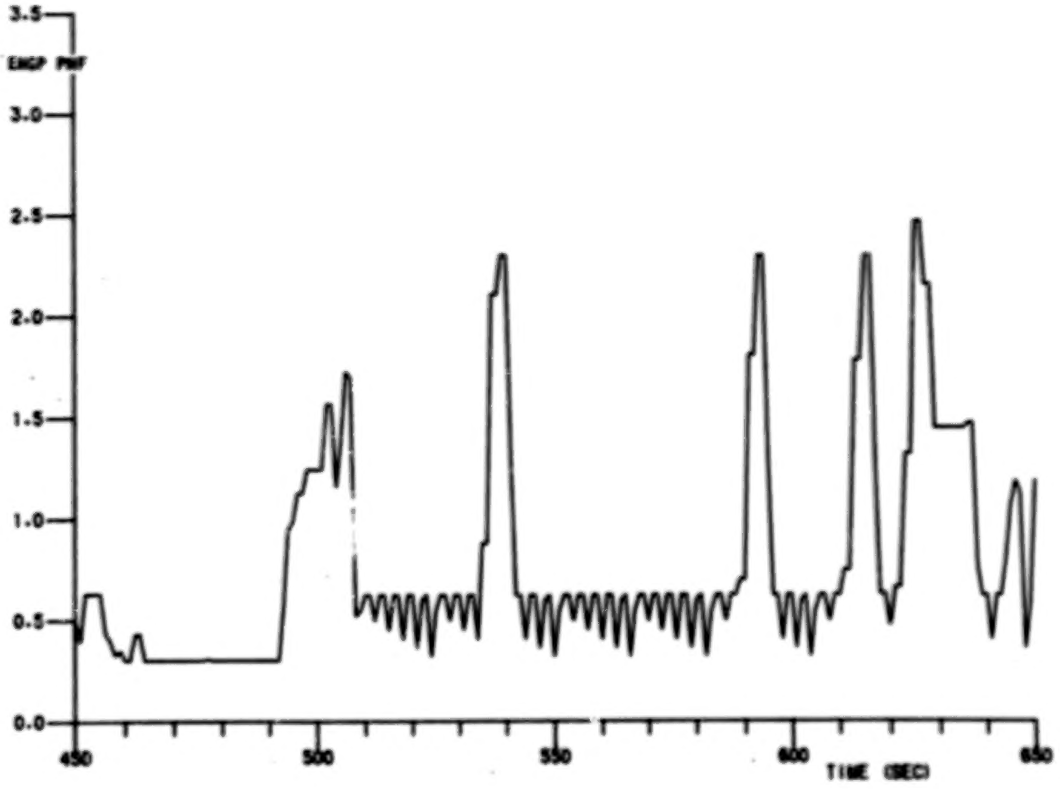


FIG. 10C SUMMATED PNF GAIN (ENGP) FUNCTIONS FOR NOMINAL APPROACH

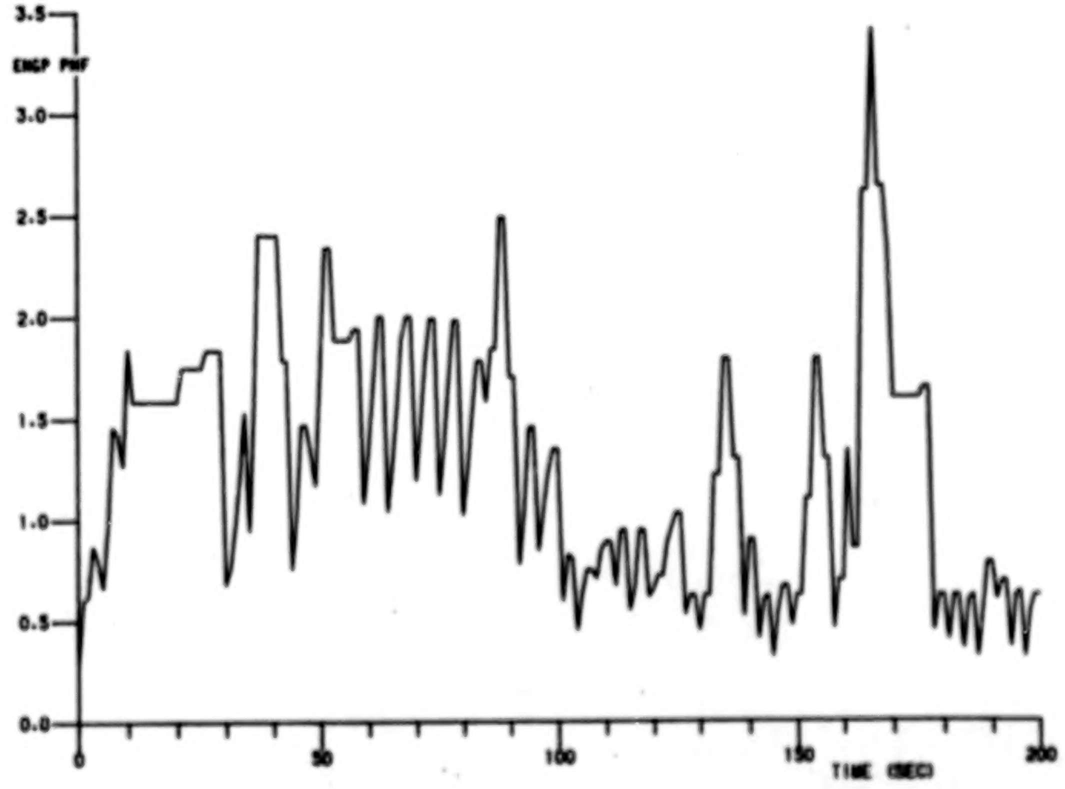


FIG. 10D SUMMATED PNF GAIN (ENGP) FUNCTIONS FOR "ACCELERATED" FINAL APPROACH

N85
14568

UNCLAS

N85 14568 D33

THE STERNBERG TASK AS A WORKLOAD METRIC IN FLIGHT HANDLING QUALITIES RESEARCH

John C. Hemingway
Ames Research Center
Moffett Field, California

SUMMARY

The objective of this research was to determine whether the Sternberg item-recognition task, employed as a secondary task measure of spare mental capacity for flight handling qualities (FHQ) simulation research, could help to differentiate between different flight-control conditions. FHQ evaluations were conducted on the Vertical Motion Simulator at Ames Research Center to investigate different primary flight-control configurations, and selected stability and control augmentation levels for helicopters engaged in low-level flight regimes. The Sternberg task was superimposed upon the primary flight-control task in a balanced experimental design. The results of parametric statistical analysis of Sternberg secondary task data failed to support the continued use of this task as a measure of pilot workload. In addition to the secondary task, subjects provided Cooper-Harper pilot ratings (CHPR) and responded to a workload questionnaire. The CHPR data also failed to provide reliable statistical discrimination between FHQ treatment conditions; some insight into the behavior of the secondary task was gained from the workload questionnaire data. A limited review of the literature on the use of the Sternberg task as a workload metric is also provided.

INTRODUCTION

The advent of sophisticated flight-control systems technologies, including digital avionics, system actuators, and fly-by-wire/fly-by-optics (FBW/FBO) systems, has lent a new dimension to the design of flight controls, and stability and control augmentation systems (SCAS). Technical innovations, including side-arm controllers, variable SCAS, and advanced multimode displays, are commonplace in modern military and commercial aircraft. Previous constraints imposed on control system designs by conventional mechanical, hydraulic, and electromechanical flight-control systems no longer apply. The contemporary design engineer must approach flight-control systems design within the total mission/systems context. He is more likely to be constrained by his own creative abilities than by available technologies. Greater operational flexibility, if not improved cost, reliability and maintainability figures-of-merit ensure the continuing application of modern technology. These advanced designs are resulting in greater demands on flight handling qualities (FHQ) research personnel to glean the full advantage afforded by the new technologies, and to determine optimal man-machine mixes.

Historically, FHQ research has relied on the abilities of highly trained and experienced test pilots to assess the adequacy of handling qualities in developmental aircraft. However, FHQ rating scales, such as the Cooper-Harper pilot rating (CHPR) scale, only generally reflect system performance and pilot workload; moreover, they are subject to pilot biases (ref. 1). Traditional FHQ rating scales alone may not be sufficiently sensitive to provide adequate discrimination between the sophisticated flight-control systems of contemporary aircraft. Thus, future FHQ research may be faced with an important problem of finding a sufficiently sensitive metric

PRECEDING PAGE BLANK NOT FILMED

to permit system performance distinctions to be made, particularly across the central and negative portions of the FHQ scale (ref. 2).

One possible solution being investigated by several researchers is to assess the pilot's spare mental capacity while performing a control task to obtain an index of the workload associated with a specific control system and, by extension, the FHQ of the system under study. In the current investigation, a relatively straightforward secondary task — the Sternberg task — was superimposed upon the pilot's primary flight-control task in order to assess reserve mental capacity. This was an exploratory investigation in which the primary objective was to determine whether performance on the Sternberg item-recognition task can provide an objective index of pilot mental workload in an FHQ simulation study.

BACKGROUND

The Sternberg task evolved from Donders' early work (ref. 3) using subtraction methodology to measure component processing times for stages believed to exist between stimulus onset and response execution. A translation of Donders' work is given in Koster (ref. 4) together with an extension of the methodology by Sternberg into an additive-factor method for detecting processing stages, assessing their attributes, and for determining their stochastic independence.

A description of the basic task along with findings from two exploratory investigations on human memory scanning was provided by Sternberg (ref. 5). A typical stimulus ensemble for the Sternberg task consists of a homogeneous set of elements (K) from which n elements are randomly drawn to compose a positive set. The remaining (K-n) elements comprise the negative set. Before each set of trials, test subjects are required to memorize the positive set, which typically varies in size from one to six elements. Stimuli are presented serially, randomly drawn from either the negative or positive subsets of K. The subject's task is to perceive a displayed stimulus, decide whether it is from the positive or negative set, and respond appropriately as rapidly as possible within the fixed interstimulus interval (ISI). Correct responses and reaction-time (RT) data from stimulus onset to response execution are collected for analysis. Response error rates for the Sternberg task are normally between 1% and 2%. Sternberg RT data versus memory set sizes are commonly presented as linear functions via regression analyses. The y-intercept value is interpreted as time for stimulus processing and response formulation, independent of set size; the slope is interpreted as the rate of search through short-term memory.

This RT measurement methodology was originally proposed by Sternberg (ref. 5) for studying the retrieval of symbolic information in short-term memory, which he later extended to research into the mechanisms underlying human information processing. In a more recent paper, Sternberg (ref. 6) reexamined the assumption underlying the basic item-recognition paradigm, and discussed the implications of findings of other researchers for the methodology and model. Four of the findings most relevant to this research are (1) mean RTs increased linearly with the size of the memory set; (2) negative and positive responses produced approximately the same slope; (3) the rate of increase is approximately 38 msec for each additional element of the positive set; and (4) the y-intercept varies about a central value of 400 msec. These basic properties have been reaffirmed by many researchers and have been interpreted to mean that memory search is "exhaustive," or sequentially completed for all elements in a given memory set.

In addition to the value of Sternberg methodology in studying basic mechanisms of information processing (see refs. 7 and 8), the relative stability of performance on the memory search task makes it an attractive candidate as a secondary task for studying workload. Knowles (ref. 9) stated that a secondary or auxiliary task can be used to discover how much additional work an operator can undertake while still performing the primary task to some specified system criteria; Knowles and Rose (ref. 10) indicated that secondary task performance is sensitive to differences in problem difficulty; that it reflects increased ease in handling the control task with practice; that it reflects differences in workload between crewmembers; and that it exposes control law difficulties during critical flight segments.

Researchers at the University of Illinois (refs. 8 and 11) have been using variations of the item-recognition paradigm in basic information processing studies to investigate structural, capacity, and resource theories of attention, including dual-task performance. Central to much of this work is a structural model of information processing based on a multiple resource theory of attention (ref. 12) which has implications for workload measurement and task integration. The model presupposes the structure of resources to include processing stages, processing modalities, and processing codes. Examples of processing stages included perceptual and central processing, response selection, and execution. Modalities included visual and auditory inputs and manual and vocal responses. Codes could be either verbal or spatial.

Wickens et al. (ref. 13) evaluated the model's ability to define resource reservoirs in a series of experiments that employed the Sternberg task among others. A primary compensatory tracking task with either rate or acceleration control dynamics was administered in the first experiment. A Sternberg task variant was administered as a concurrent task on selected trials. Dimensions that were varied on the secondary task included perceptual load (superimposing a visual grating), central processing load (memory set size), and response load (single- versus double-key press). The results of this investigation showed that stage-defined resources could be successfully differentiated by the Sternberg task. One surprising result was that RT's were not affected when the double response load was superimposed upon the secondary task. The authors suggested that the additional load was reflected in poorer performance on the primary tracking task.

In a second experiment by Wickens et al. (ref. 13), a failure-detection task was paired with the Sternberg task to evaluate the function of spatial and verbal processes in defining resource reservoirs for encoding and central processing stages. Contrary to predictions based on multiple resource models of attention, differences in slope for longer memory set sizes for both verbal and spatial stimuli failed to materialize. Instead, differences between single- and dual-task performance were reflected in the y-intercept values, which were elevated particularly for the spatial condition. The authors pointed out that higher intercept values were consistent with multiple resource theory, placing greater demands on spatial instead of on verbal resources for perceptual interactions between the central processing loads (memory set sizes) imposed on the Sternberg task.

In a follow-on study, Wickens et al. (ref. 14) conducted three experiments to examine coding effects on performance. Baseline data on verbal and spatial Sternberg tasks were collected in the first experiment. Data from both variants produced generally linear functions, although the authors reported finding a significantly greater slope for the spatial condition, as well as a weak quadratic tendency in the function. Reliable interactions were reported between memory set size, verbal versus spatial stimuli, and response hand, which the authors interpreted as providing evidence for separation and resource competition between and within hemispheres for processing these stimuli. The same Sternberg task variants and stimuli were used in the second experiment in combination with a memorization task. Results suggested that the spatial secondary task shared more common resources with the memory task than its verbal counterpart. The authors concluded that the spatial/verbal dichotomy is an important element in interpreting dual task interference patterns.

The results of the third experiment, which shared an autopilot monitoring/failure-detection task with the same two Sternberg tasks, indicated that more interference occurred with the spatial variant than with the verbal, because the failure-detection task was spatial in nature. As in the previous investigation by Wickens et al. (ref. 13), no interaction was discovered between dual-task load and memory set size. Rather than indicating that the primary task had no perceptual or central processing demands, the additivity seemed to be related more to the automaticity of certain processes. Theoretical considerations surrounding the use of the Sternberg additive factor method for assessing the demands of the primary task have been discussed within a multiple-resource modeling context in several other publications (refs. 11, 12, and 15).

The simplicity and relative stability of the Sternberg task methodology make it suitable for multimodality research as well. Vidulich and Wickens (ref. 16) reviewed recent multimodality research, and reported findings from two experiments in which Sternberg tasks were used to investigate differences between combinations of input (auditory or visual) and output (verbal or manual) modalities. Findings from these investigations are described in terms of multiple-resource theory, and help to clarify both code and modality relationships in facilitating time-sharing efficiency. Two of their findings (ref. 16) of particular relevance to the current investigation are:

1. Task priorities exerted a reliable effect on performance, and this effect was greater as the tasks shared more common resources.

2. Although clear performance differences were observed between input/output modality conditions, these were not reflected in the assessment of subjective workload ratings.

Observer ratings were not sufficiently sensitive to judge dual-task demands; however, the demands of different types of primary tasks could be assessed by specific types of secondary tasks, with greatest sensitivities obtained when task demands tap common resource pools. In other words, sharing input modalities may lead to a deterioration of the intermittent discrete reaction time task, whereas sharing outputs could lead to a deterioration of the continuous manual task (ref. 17).

Several researchers have employed Sternberg task methods to evaluate workload demands in real and simulated flight. Crawford et al., (ref. 18) using a cockpit simulator, evaluated two levels of flight control and four levels of multifunction switching, using the Sternberg task as a secondary measure of reserve information-processing capacity. Performance on the Sternberg task differed by 54% between flight-control levels, and by 20% to 31%, respectively, for simple and complex multifunction switching tasks. Corrick (ref. 19) employed the Sternberg task to evaluate four alternative display formats used to present missile launch envelope information. Although subjects failed to report any large performance-related differences between the displays, the author found large differences in secondary task performance which they attributed to the workload imposed by display formatting of the primary task.

An interesting variation of the Sternberg task was employed by Johnson (ref. 20) in studying the effects on reaction time of terrain background, downlook angle, and response-processing levels in target acquisition. The stimulus ensemble consisted of eight tank targets in place of traditional alphanumeric stimuli. Three groups of five subjects each were required to make a positive or negative set determination, recognize a target (friend or foe), or identify the target. Reaction-time data were collected and analyzed for memory set sizes of one, two, and four for each of the three acquisition tasks. Greatly inflated y-intercept (1,400 vs 400 msec), and slope (200 vs 40 msec per memory set size) values over those reported by Sternberg (ref. 6) were attributed to differences in target and task complexities. Statistically significant performance differences were found between levels of target background, downlook angles, and acquisition tasks. The results were consistent with Sternberg's findings, supporting the serial exhaustive model of memory search, and the authors concluded that the Sternberg task method served as a useful tool in understanding the observer's cognitive processes in complex target acquisition tasks.

Schiflett et al. (ref. 21) used the Sternberg task in actual flight tests on board a T-33 variable-stability research aircraft. The goal of the study was to evaluate two levels of flight control and two alternative head-up display (HUD) formats under simulated instrument meteorological conditions. The primary control task flown in the T-33 fixed-wing aircraft consisted of glide-slope and localizer intercepts, and an ILS approach to touchdown. Subjects flew four approaches for each combination of display conditions and handling-quality level. The findings of Schiflett et al. were similar to those of Corrick (ref. 19) — poor agreement was obtained between the primary flight performance measures, Cooper-Harper ratings across primary flight-control task configurations, and performance on the secondary task. It appeared that the pilots compensated for the more demanding flight-control variations, but at the expense of reserve information-processing capacity. Of particular interest were the apparent sensitivities of measures obtained on the secondary task for positive memory set sizes of one, two, and four. All data appeared to fit the Sternberg paradigm remarkably well, with both slope and y-intercept discriminants showing consistency across levels of flight-control and display type for both subjects. It should be noted, however, that no parametric statistical analyses were performed, and that analysis was restricted to exploring trends in reaction-time and response-error data. The authors concluded that the pilots had more mental reserve capacity while flying the predominantly pictorial/symbolic HUD configuration than when flying the conventional HUD format with scales and alphanumerics; however, they recommended further research be done to establish the efficacy of the Sternberg task in evaluating aircrew tasks.

Pretest and Findings

In order to determine if the Sternberg task could be effectively integrated into an FHQ simulation study, a preliminary investigation was performed in a fixed-base simulator at Ames Research Center. The Sternberg task was superimposed on a primary flight-control task for two pilot subjects during a comparative evaluation of an integrated isometric controller for nap-of-the-earth (NOE) flight (ref. 22). In addition to the secondary task, measures of pilot control activity, CHPR data, and pilot commentary were collected across the different experimental conditions. The experimental protocol for the Sternberg task was previously described; the methodology approximated that described by Schiflett (ref. 23) employing the same alpha stimulus ensemble, the same positive set sizes (one, two, and four), and the same 7.0-sec interstimulus interval (ISI). The primary flight-control task was performed in a fixed-base, rotary-wing simulator, and typified an NOE mission with maneuvering, hover and bob-up, and straight-and-level flight segments.

Linear regression fits of the reaction-time data obtained on the baseline condition (secondary task alone) generally conformed to the classical Sternberg model for serial probe recognition. The y-intercepts for the two subjects, 665 and 598 msec, respectively, were longer than the typical 400-msec values reported by Sternberg.[1] Slope values for ascending memory set sizes were 43 and 27 msec, respectively, compared with the 38-msec value reported by Sternberg (ref. 6). Numbers of reversal and time-out errors ($RT > 500$ msec) on the baseline task were less than 1%.

The behavior of the Sternberg function became erratic for both subjects with the addition of the primary flight-control task. Specifically, y-intercept values across all experimental conditions ranged between 850 and 1,550 msec, and the slope of the function across memory set sizes ranged from a high of 47 msec per set size to a negative 111 msec. Response error rates (reversal and time-out errors) increased from 1% on the baseline condition to over 10%. RT's for these responses were discarded from other analyses.

Pilot ratings (CHPR) regarding FHQ were generally poor, ranging from a high of 3.5 (satisfactory with unpleasant characteristics) on the cruise flight segment to 7.0 (unacceptable) on the maneuvering and hover and bob-up segments. An analysis of variance of secondary task conditions showed only one statistically significant difference ($p < 0.001$) between the baseline Sternberg condition and the other combined experimental conditions, and no memory set size effects upon RT's emerged when the primary task was added. The absence of this effect was apparent from the erratic slopes obtained on the secondary task when combined with the primary task. It is important to recall, however, that data were obtained from two pilots only, and results must be viewed accordingly. Although both subjects accepted the addition of the secondary task, poor handling qualities on the primary flight-control task appeared to have inhibited their performance on the secondary task.

Although the procedural integration of the Sternberg task into an FHQ research paradigm was accomplished, it was difficult to collect sufficient numbers of data points (reaction times) for analysis without incurring primary task overload from the secondary task. Additionally, the subjects complained that the location of the secondary task response key on the collective control interfered with the manual, primary flight-control task, thus indicating an alternative response mode for the secondary task.

---

[1] A post hoc examination of the software program revealed an implementation bias which consistently inflated RT's between 70 and 100 msec. No correction was applied to these pretest data.

Research Question

The objective of the investigation was to determine whether the Sternberg task, used as a secondary task to measure pilots' spare mental capacity in FHQ simulation research, could differentiate between different flight-control systems. It was hypothesized that any change on the relatively stable Sternberg task would reflect variation in encoding, output, or processing loads of the primary task competing for common resources with the Sternberg task; an increase or decrease in the y-intercept would reflect a change in encoding or output components, while mental information processing would be reflected by changes in the slope. The secondary task input and output modalities were configured to achieve compatibility with the resource demands of the primary flight-control task. Thus, both primary and secondary tasks involved a sharing of the spatial input modality, whereas the output modality on the secondary task was switched to verbal (ref. 15) for this experiment.

Unanticipated performance decrements on the primary rather than the secondary task (ref. 13) and prioritization shifts (ref. 16) have already been discussed as potential problems in the application of this workload method. Consequently, all practice and at least one data run were completed for each FHQ experimental condition without the secondary task. If significant performance differences were subsequently found on the primary task, they would constitute evidence for rejecting the Sternberg task as a useful workload metric on future FHQ simulation research.

In addition to the aforementioned objectives, we also wanted to investigate the relationship between subjective workload assessment ratings, CHPR data, and performance on the Sternberg secondary task. Consequently, questionnaires and CHPR data were systematically collected throughout this investigation.

METHOD

This investigation was conducted in the Vertical Motion Simulator (VMS) at Ames Research Center. The VMS cab was configured to represent a generic single-seat helicopter, incorporating a conventional helicopter front instrument panel. The primary objective was to evaluate the FHQ of three different primary flight-control configurations and three stability and control augmentation levels in low-level flight regimes. A Sternberg item-recognition task was superimposed upon the primary flight-control task to determine whether this particular secondary task could provide an indication of pilot workload.

Subjects

Four helicopter test pilots served as subjects for this investigation; two were NASA test pilots; one was a Canadian National Aeronautics Establishment pilot, and one was a Boeing-Vertol test pilot. All subjects possessed current first-class medical certificates and had participated in similar helicopter FHQ research projects.

Apparatus

Flight-control system— The hardware for the three primary flight-control systems consisted of (1) a conventional helicopter cockpit configured with cyclic, collective, and yaw damper pedals; (2) an advanced configuration incorporating a four-axis, isometric side-arm controller with roll controlled by lateral force, pitch by longitudinal force, yaw by rotational moment, and thrust by vertical force; and (3) a mixed configuration similar to the four-axis controller described above, except that thrust was incorporated on the conventional collective. All flight-control hardware remained in place during this investigation, but the desired control configurations were actuated under software control (see fig. 1).

Helicopter control dynamics were generated by a SIGMA-8 computer programmed with a nine-degree-of-freedom generic teetering-rotor helicopter model (ref. 24) which included both stability and control augmentation. Details and rationale for the selection of the three levels of stability and control augmentation investigated in this simulation were described by Aiken (ref. 25). Basically they included (1) unaided: simulated basic UH-1 mechanical control system, without stabilization; (2) rate damping: augmented angular rate damping in pitch, roll, and yaw; and (3) rate-command/attitude-hold (RCAH): integral prefilters in pitch and roll to provide a rate-command/attitude-hold feature.

Visual-display system— The visual system consisted of raster- and stroke-written television monitors mounted perpendicularly to one another such that the imagery from each monitor was projected onto a common combining glass. Subjects viewed combined imagery, focused at optical infinity, through a lens system located above the front instrument panel. A realistic external visual scene, also presented to the subjects on the raster display, was generated by the Ames Visual Flight Attachment (VFA) apparatus — a gimbaled television camera mounted on a gantry system which traversed a terrain board. Camera motion was under computer control and responded appropriately in six degrees of freedom to pilot control inputs.

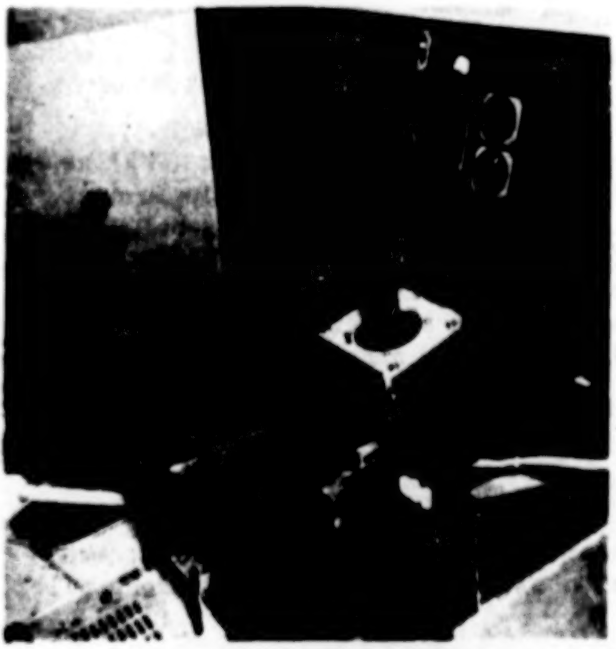


Figure 1.- Flight-control hardware.

The stroke-written television monitor contained all critical flight parameters, advisory information, and Sternberg secondary task stimuli. All Sternberg task display characters exceeded the threshold for limiting resolution and were presented just to the left of central field-of-view. Display-mode select control functions were interfaced with a PDP-11/55 computer that generated the stroke-written symbols. Primary flight display symbology consisted of modified pilot night vision system

(PNVS) display formats for various flight modes, including descent, accelerating and decelerating transitions, hover and bob-up, and cruise (ref. 26); see fig. 2.

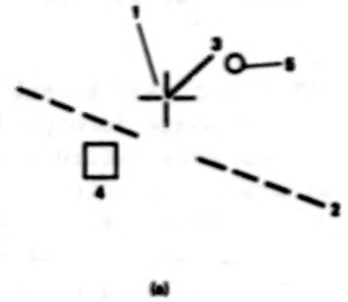
Mission scenario— A predefined mission profile comprising four distinct flight segments was used throughout this investigation. Figure 3 shows the mission segments, including a 457-m (1,500-ft) AGL, 6° descent into a nap-of-the-Earth (NOE) ground track composed of maneuvering, hover and bob-up, and cruise flight segments. The profile was defined such that each flight segment required approximately 2 min to complete.

Data recording— A Voterm voice recognition system (VRS) was used by subjects in responding to the Sternberg secondary task to avoid manual response incompatibility problems identified in the preliminary study. Reaction times were recorded by calculating elapsed time from stimulus onset to the first utterance detected by the VRS. Subjects' digitized responses (yes or no) and RT data were stored on the simulation computer following each run, together with handling qualities data. In addition to FHQ objective performance measures, Cooper-Harper ratings, pilot comments concerning each of the four mission segments, and responses to a postflight questionnaire were obtained at the end of each simulation run. Each subject also completed a final debriefing questionnaire following the simulation.

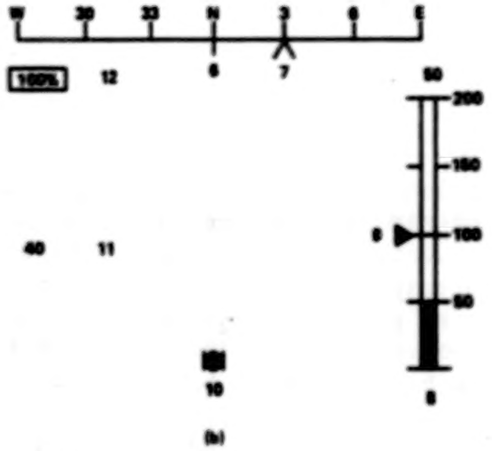
Procedure

System performance and calibration checks, including picture quality assessments and flight-control response dynamics, were carried out before each test session. The Sternberg secondary task procedure was identical to that described for the pretest, except that subjects were required to respond aurally, instead of manually. The four subjects were acquainted with the primary flight-control task, introduced to the Sternberg secondary task, and given a pre-recorded set of instructions that explained the procedures and performance priorities to be employed on the primary and secondary tasks. They were told that they would be required to fly an 8-min NOE mission composed of a 1,500-ft, 6° descent; low-altitude maneuvering; hover and bob-up; and straight-and-level flight segments. The subjects were also furnished with the appropriate altitude and airspeed requirements for each segment. The PNVS display-mode switching control functions were explained for each mission segment. The subjects were instructed to maintain a high and constant level of performance on the primary flight-control task, although their performance would be scored on the Sternberg secondary task as well. They were reminded that although the letters from the secondary task would remain on the display for 5 sec, they should respond as rapidly as possible after making their yes or no decision. Following each run, subjects were required to assign a Cooper-Harper rating to each of the four mission segments and to provide comments on selected facets of both the primary and secondary tasks.

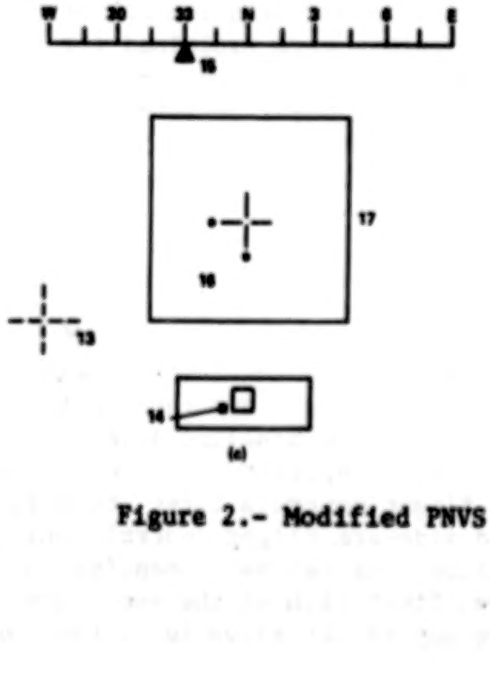
After receiving their instructions, the subjects were brought into the simulation cab, given several practice sessions on the Sternberg task, and then tested solely on the secondary task to ensure proficiency and obtain baseline data. The core experiment, comprising the flight-control conditions depicted in table 1, consisted of testing all subjects on the conventional flight controls first, then running a pair of subjects through one of the advanced side-arm flight-control configurations. All subjects were permitted familiarization runs before commencing data collection on each primary flight-control condition, first without the secondary task, then with balanced presentations of the three memory set sizes (one, two, and



SYMBOL	INFORMATION
1. Aircraft reference	Fixed reference for horizon line, velocity vector, hover position, cyclic director, and fire control symbols
2. Horizon line (cruise mode only)	Pitch and roll attitude with respect to aircraft reference (indicating nose-up pitch and left roll)
3. Velocity vector	Horizontal Doppler velocity components (indicating forward and right drift velocities)
4. Hover position	Designated hover position with respect to aircraft reference symbol (indicating aircraft forward and to right of desired hover position)
5. Cyclic director	Cyclic stick command with respect to hover position symbol (indicating left and aft cyclic stick required to return to designated hover position)



SYMBOL	INFORMATION
6. Aircraft heading	Moving tape indication of heading (indicating North)
7. Heading error	Heading at time bob-up mode selected (indicating 030)
8. Radar altitude	Height above ground level in both analog and digital form (indicating 50 ft)
9. Rate of climb	Moving pointer with full-scale deflection of ±1,000 ft/min (indicating 0 ft/min)
10. Lateral acceleration	Inclinometer indication of side force
11. Airspeed	Digital readout in knots
12. Torque	Engine torque in percent



SYMBOL	INFORMATION
13. Cued line of sight	Overlays designated target position on background video when target is in display field of view
14. Coarse target location	Designated target position with respect to display field of view (inner rectangle) and sensor limits (outer rectangle)
15. Target bearing	Designated target bearing (indicating 330° or 30° to left of current heading)
16. Target location dots	Illumination of two adjacent dots indicates display quadrant in which designated target is located
17. Missile launch constraints	Limits with respect to aircraft reference for successful weapon lock-on to designated target

Figure 2.- Modified PNVS display symbols (from Aiken, ref. 25).

ORIGINAL PAGE IS
OF POOR QUALITY

ORIGINAL PAGE IS
OF POOR QUALITY

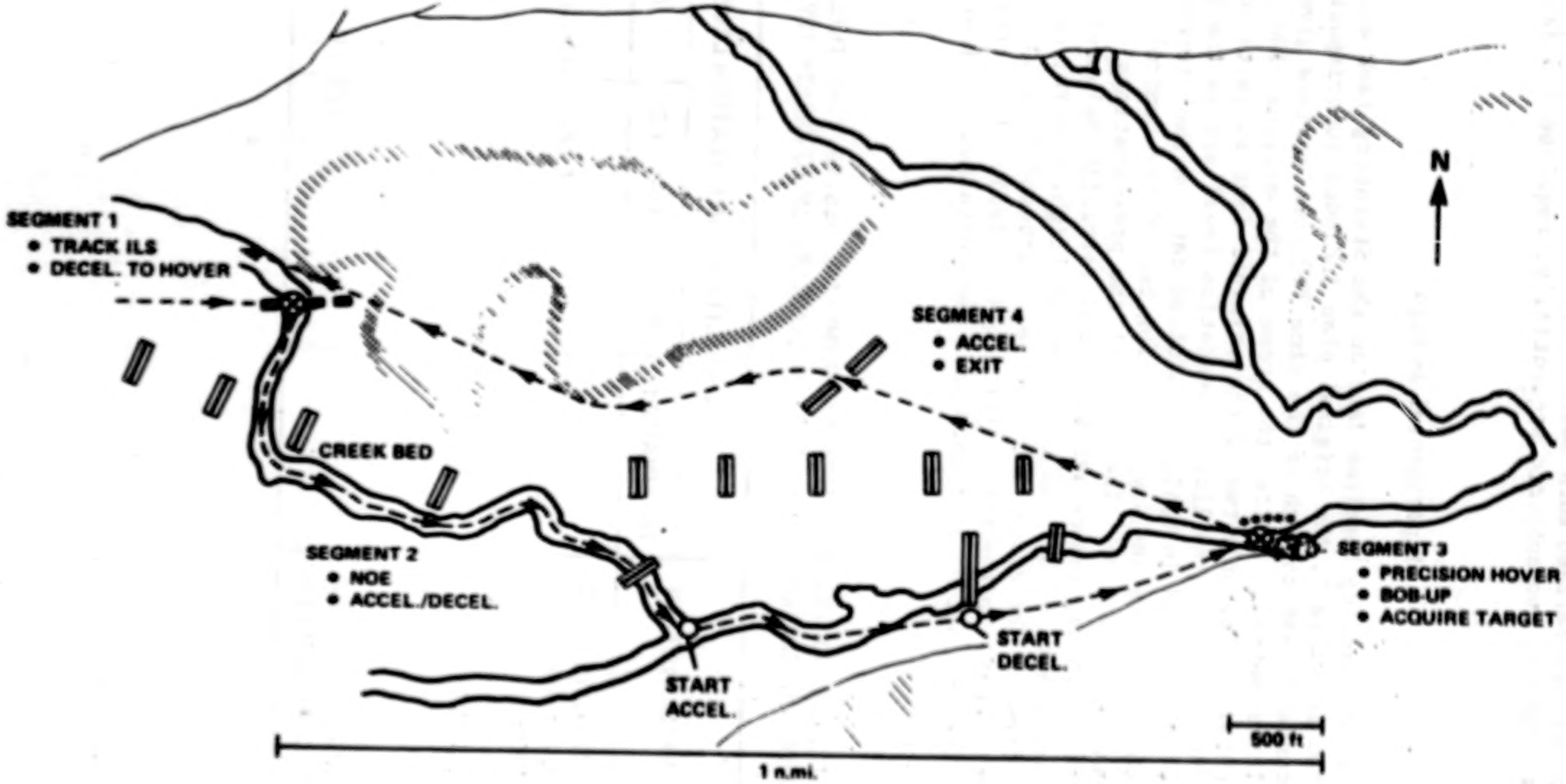

SEGMENT 1
• TRACK ILS
• DECEL. TO HOVER
CREEK BED
SEGMENT 2
• NOE
• ACCEL./DECEL.
START ACCEL.
START DECEL.
SEGMENT 3
• PRECISION HOVER
• BOB-UP
• ACQUIRE TARGET
SEGMENT 4
• ACCEL.
• EXIT
N
500 ft
1 n.mi.

Figure 3.- Predefined mission segments for primary flight-control task.

four) of the Sternberg task while flying the mission segments. The cab controls were then reconfigured, and the subjects were tested on the other advanced control configuration. The other pair of subjects received the same treatment, except in reverse ordering of the advanced side-arm flight-control configurations. Cooper-Harper ratings and pilot comments were solicited after each run, and a 15-mi rest period was permitted between each control-system reconfiguration. Training and test sessions, running two subjects sequentially, required 1.5 days to complete per pilot.

Experimental Design

In addition to collecting baseline data on the Sternberg task and on the FHQ conditions depicted in table 1, the original plan included superimposing three levels of the secondary task on each of the nine FHQ combinations given in table 1. Computer difficulties severely limited the scope of the original FHQ design; complete data were obtained for only two of the original four subjects for the rate-command/attitude-hold (RCAH) stability augmentation level across the three control configurations as illustrated in table 1. Limited data on two Sternberg memory set sizes were also obtained from one subject flying two of the control configurations under the rate-damping condition. The order of the presentations of control configurations and of Sternberg memory set sizes was initially balanced across four subjects, with the exception that the conventional flight-control configuration was always presented first. The design was to have progressed from the best stability augmentation level (RCAH) to the most difficult (unaided). The actual order of presentation for the two subjects on whom data were obtained is shown in table 2.

TABLE 1.- FLIGHT HANDLING QUALITIES EXPERIMENTAL DESIGN (THE THREE STABILITY AUGMENTATION LEVELS WERE TO HAVE BEEN EXAMINED IN SEPARATE INVESTIGATIONS

CONTROL CONFIGURATION	STABILITY AUGMENTATION LEVEL		
	(A) UNAIDED	(B) RATE DAMPED	(C) RCAH
(1) CONVENTIONAL (CYCLIC, COLLECTIVE, YAW DAMPER PEDALS)		▨ EXPERIMENTAL	▨
(2) THREE AXIS FORCE STICK AND COLLECTIVE (SIDE-ARM CONTROLLER)			▨ CONDITIONS
(3) FOUR AXIS FORCE STICK (SIDE-ARM CONTROLLER)		▨ TESTED	▨

TABLE 2.- ACTUAL PRESENTATION ORDER OF SECONDARY TASK MEMORY SET SIZES WITHIN THE FHQ EXPERIMENTAL DESIGN

SUBJECT	EXPERIMENTAL CONDITIONS	SESSION NUMBER (4 RUNS/SESSION)				
		1	2	3	4	5
3	SCAS/CONTROLLER*	C1	C2	C3	B1	B3
	MEMORY SET SIZE	0, 1, 2, 4	0, 4, 1, 2	0, 2, 4, 1	0, 4, 1	0, 1, 4
4	SCAS/CONTROLLER*	C1	C2	C3		
	MEMORY SET SIZE	0, 2, 4, 1	0, 4, 1, 2	0, 1, 2, 4		

*SEE TABLE I

RESULTS AND DISCUSSION

FHQ aspects of this research project, including the pilots' evaluations of the adequacy of the three flight-control configurations and stability augmentation levels (table 1) for helicopter terrain flight, were reported by Aiken (ref. 25, p. 5). Aiken presented the following conclusions:

> 1. With conventional controllers, a rate- or attitude-stabilized vehicle, and a head-up display, adequate but unsatisfactory handling qualities were achieved for the low-altitude tasks investigated.
>
> 2. Satisfactory handling qualities may be achieved with a head-up display and a properly designed two-axis displacement side-stick controller for either a rate- or attitude-stabilized vehicle. Critical side-stick design features include the force-deflection characteristics and mechanization of the trimming function.
>
> 3. Attitude stabilization is required to maintain adequate handling qualities with either the rigid three-axis (pitch, roll, and yaw) or four-axis controller configuration evaluated during this investigation.

The remainder of this report addresses workload-related findings using the Sternberg task, and the effect of introducing a secondary task on the primary flight-control task.

Linear regression fits of the Sternberg baseline data obtained from the four subjects in this experiment are shown in figure 4. These data were corrected for length of utterance and Voterm processing times to enable comparison with classical Sternberg task results. The curves and means were adjusted downward for subjects 1 and 2 by averaged yes and no utterance times of 530 and 540 msec, respectively, plus the 224-msec Voterm processing time. For subjects 3 and 4, individual utterance and processing times were recorded and subtracted on each experimental trial. There were

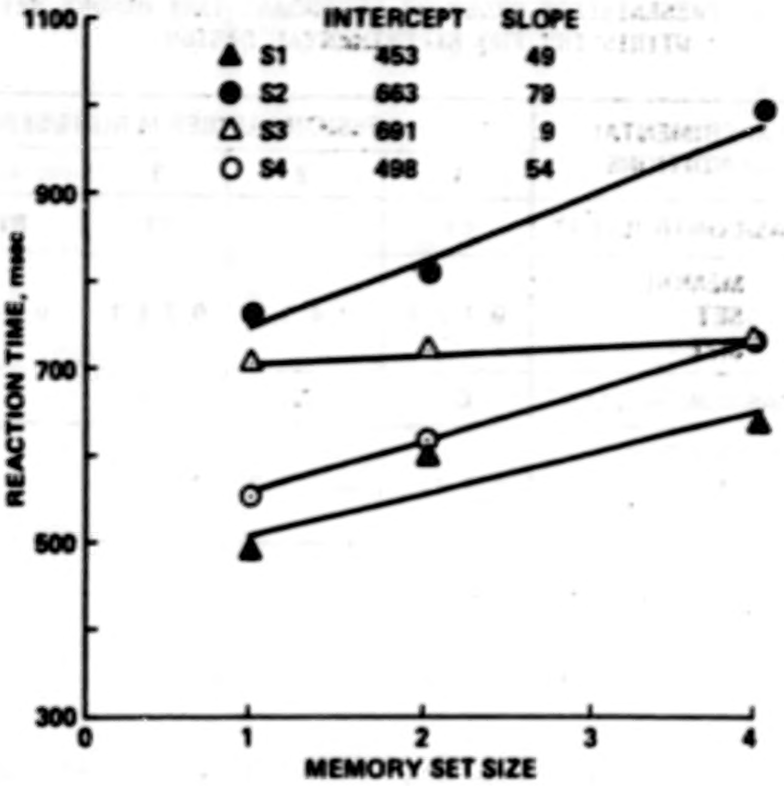


Figure 4.- Plots of subjects' baseline Sternberg data. Symbols represent means for each subject for memory set sizes one, two, and four. Error rates were insignificant over these data.

24 trials per memory set size; however, only 15 of the original 24 discrete RT's could be used per cell for baseline data analyses — a result of unrecoverable data on subject 3. Consequently, the last nine trials were deleted from each cell of the remaining three subjects to facilitate analysis. Error rates (reversal errors) for all subjects remained relatively constant, varying between 1% and 2% — results similar to those reported by Sternberg (ref. 5).

The results of an ANOVA (ref. 27) performed on the baseline data for the four subjects are given in table 3. Both main effects, subjects (S), and memory set size (M) were highly significant, ($F = 21.45$ ($df = 3,42$) $p < 0.0001$) and ($F = 19.54$ ($df = 2,28$) $p < 0.0001$), respectively, and there were no significant interactions. A strength-of-association statistic — omega-squared (ref. 28) — was calculated for both statistically significant main effects. Subjects accounted for about 27% of the RT variance, and memory set size captured about 12%.

Four ANOVAS (ref. 27) were calculated using each subject's baseline Sternberg data to separate out the effects of memory set size on RT. A summary of these analyses together with r-squared strength-of-association measures is shown in table 4. The results showed that memory set size had a significant effect on the performance of all but the third subject. The r-squared values showed that memory set size captured about 9%, 22%, 0.8%, and 20% of the RT variance of the four subjects, respectively.

An orthogonal polynomial regression (ref. 27) was run on these data to investigate the linearity assumption of the Sternberg model. The results showed that the

TABLE 3.- RESULTS OF ANOVA (ref. 27) PERFORMED ON STERNBERG BASELINE DATA; STRENGTH OF ASSOCIATION VALUES (ref. 28) INDICATE TOTAL ACCOUNTED VARIANCE

Source	df	Sum of squares	Mean squares	F	p	Strength of association
Subjects	3	2141213.75	713737.92	21.45	0.0001	0.2680
Error	42	1397803.50	38288.39			
Memory set (M)	2	922212.31	461106.16	19.54	.0001	.1154
Error	28	660737.86	23597.78			
SM	6	210458.80	35076.47	1.39	N.S.[a]	--
Error	84	2122299.70	25265.47			
Mean	1	83527381.61	83527381.61			--
Error	14	536037.48	38288.39			

[a] not significant.

TABLE 4.- RESULTS OF SEPARATE ANOVA's (ref. 27) PERFORMED ON MEMORY SET SIZE FOR ALL SUBJECTS; THE PERCENT OF VARIANCE CAPTURED BY MEMORY SET SIZE IS REFLECTED IN THE r-SQUARED STRENGTH OF ASSOCIATION MEASURES

Source	df	Sum of squares	Mean squares	F	p	Strength of association
Memory set ($\underline{S}_1$)	1	196984.75	196984.75	4.622	0.05	0.0862
Error	49	2088382.00	42620.04			
Memory set ($\underline{S}_2$)	1	510218.69	510218.69	13.961	.001	.2182
Error	50	1828596.00	36571.92			
Memory set ($\underline{S}_3$)	1	10259.06	10259.06	.643	N.S.	.0081
Error	79	1259894.00	15948.02			
Memory set ($\underline{S}_4$)	1	357086.87	357086.87	18.369	.001	.1967
Error	75	1457935.00	19439.13			

slope was significant for all but the third subject, and that no significant quadratic term existed for any of the data sets. The polynomial regression analysis is summarized in table 5.

As previously discussed, all secondary task data for subjects 1 and 2 under dual-task experimental conditions were unrecoverable because of data acquisition problems. All remaining RT data for subjects 3 and 4 on the secondary task were summarized by flight-control configuration, stability augmentation level, and mission segment (table 6). In the original design, 2 min of flight time were planned for each of the four mission segments in order to ensure near-equal data samples across segments. In actuality, pilots varied greatly in the amount of time they devoted to each segment, presumably in the interests of performing FHQ evaluations. Thus, there were unequal numbers of data points obtained for each of the three

TABLE 5.- RESULTS OF POLYNOMIAL REGRESSION ANALYSIS (ref. 27) RUN ON SUBJECTS' BASELINE RT DATA FOR THREE MEMORY SET SIZES; ANALYSIS PROGRAM AUTOMATICALLY TERMINATES AT NEXT HIGHEST VALUE

Subject	Degree	Multiple r-squared	Regression coefficient	df	F	p
1	0	-	4776.4954	2	3.04	0.0572
	1	0.08619	443.8203	1	1.51	.2250
	2	.11467	-255.1048	-	-	-
2	0	-	7741.5346	2	6.90	.0023
	1	.21815	714.2853	1	.31	.5793
	2	.22319	108.5881	-	-	-
3	0	-	6416.0000	2	.36	.6988
	1	.00807	101.2619	1	.09	.7616
	2	.00927	-38.9191	-	-	-
4	0	-	5483.3271	2	8.95	.0003
	1	.19673	597.5495	1	.02	.8970
	2	.19691	-18.3496	-	-	-

memory set sizes. The smallest sample size was six data points per set size; the largest was 44 data points per set size. Pilots also failed to respond to trials (out-of-time errors), particularly during the descent and hover and bob-up mission segments. Failure to respond during the descent was the logical consequence of requiring pilots to perform a heads-up secondary task while flying a head-down ILS approach to breakout at 100 ft AGL. Failure to respond during hover and bob-up occurred because pilots attended to external visual cues, rather than to their primary, head-up flight display. Fifty-one out-of-time errors were recorded on the core (RCAH) experiment, but were not included in the data analysis. Reversal errors (yes/no confusions) were between 1% and 2% in the baseline condition.

Reaction times, and presumably pilot attention, varied greatly in the presence of the primary flight-control task. In order to improve normality or equality of variance, several possible transformations were considered, including logarithmic transforms and Windsor transforms (ref. 29). Logarithmic transforms of these data tended to reduce the problem of outliers, not atypical of such data, accompanied by a dramatic decrease in standard deviations. Unfortunately, this transform also obscured relatively large differences in slope and intercept values, and prevented interpretation of the data within the context of the Sternberg paradigm.

In addition to these manipulations, subjects' RT data were replotted using a technique proposed by Schiflett (ref. 23) which involves a truncation of the secondary task RT's above a predetermined value (1500 msec). The application of the technique discussed by Schiflett, using an arbitrary 1500-msec RT cutoff, tended to normalize many of the extreme excursions in the data sets. The following apparent benefits were noted:

1. By truncation of data sets at 1500 msec, y-intercept values, artificially inflated or deflated by outlying RT data, were decreased or increased, respectively. This also facilitated comparisons with the Sternberg task baseline data.

ORIGINAL PAGE IS
OF POOR QUALITY

TABLE 6.- MEMORY SET SIZE MEANS, STANDARD DEVIATIONS, SLOPES, AND INTERCEPTS. FROM THE STERNBERG TASK RT DATA SHOWN AS A FUNCTION OF THE FLIGHT CONTROLLER, SCAS, AND FLIGHT SEGMENT (see fig. 3)

Subject number	Experimental conditions	SEG	Memory set size X(σ) 1	2	4	Slope	Intercept
3	Baseline		697(98)	716(152)	726(114)	9.18	692
	RCAH/Conv	D	1169(619)	975(515)	1151(550)	1.25	1094
		N	873(401)	1256(1136)	1464(748)	182.71	774
		H	910(455)	1097(1127)	1051(448)	34.85	936
		S	651(167)	1138(936)	1236(692)	172.97	603
	RCAH/3-Ax	D	970(351)	1895(1350)	1425(913)	76.00	1281
		N	1246(881)	1263(692)	1217(431)	-11.98	1270
		H	1049(418)	1217(715)	1064(338)	-5.60	1144
		S	1293(478)	958(453)	978(348)	-89.86	1297
	RCAH/4Ax	D	1057(571)	1513(935)	1726(1116)	202.92	961
		N	1027(493)	1109(459)	1088(598)	13.79	1044
		H	1034(411)	1356(939)	1056(740)	-9.48	1162
		S	749(172)	1044(430)	941(197)	49.12	806
	RD/Conv	D	1179(860)	-	1388(1011)	69.70	1109
		N	1157(539)	-	907(289)	-88.33	1241
		H	1385(1059)	-	1022(192)	-120.70	1505
		S	862(415)	-	910(344)	16.13	845
	RD/3-Ax	D	1485(740)	-	1504(875)	6.32	1479
		N	1023(456)	-	1007(253)	-6.95	1030
		H	1145(441)	-	1100(488)	-14.92	1160
		S	1246(689)	-	1149(760)	-32.34	1278
4	Baseline		550(108)	610(167)	714(133)	54.27	498
	RCAH/Conv	D	738(243)	989(417)	1449(775)	236.32	507
		N	870(315)	790(149)	1209(603)	126.95	663
		H	674(110)	1010(317)	1119(215)	130.61	635
		S	778(139)	717(151)	922(367)	56.93	672
	RCAH/3-Ax	D	822(162)	1844(1122)	1862(865)	263.61	957
		N	1199(759)	1274(898)	1115(472)	-37.94	1287
		H	1160(936)	1205(447)	1379(948)	76.50	1068
		S	797(215)	1105(484)	968(324)	33.24	883
	RCAH/4-Ax	D	2232(1492)	1561(550)	1396(911)	-229.23	2253
		N	884(468)	935(249)	1127(445)	82.24	791
		H	1927(1310)	1311(890)	1599(1248)	-48.72	1668
		S	957(460)	-	879(126)	-32.90	1016
	RCAH/3-Ax	D	1104(666)	-	1556(956)	150.95	953
		N	930(267)	-	965(312)	11.48	919
		H	992(432)	-	1450(1039)	152.72	839
		S	1159(284)	-	1158(347)	-.31	1159

2. Radically high, as well as negative, slope values were decreased, or made positive by this truncation, with only three exceptions.

3. Standard deviations were dramatically and consistently decreased across all set sizes using truncated data sets.

Despite apparent advantages derived from this treatment, adequate justification for discarding outlying data was not available in this investigation because subjects were not informed of a predetermined cutoff for acceptable RT's. In fact, numerous latencies greater than 1500 msec, believed to be the result of task overload, were noted on secondary task responses during the actual simulation, despite instructions to respond as rapidly as possible. This finding is not only relevant to the feasibility of establishing empirical criteria for defining an out-of-bounds type error, as discussed by Schiflett (ref. 23), but also affects the potential utility and interpretability of secondary RT task data within the Sternberg paradigm. The possible application of Windsor transformations to these limited sets of RT data was rejected for reasons similar to those stated above, but might be considered as an alternative to truncating actual data when greater numbers of observations can be taken within selected set sizes.

The RT data obtained from this investigation do not fit within the classical Sternberg interpretation. Additionally, transformations and manipulations of the data set failed to improve interpretability within this context. It was not evident whether this was a result of (1) primary-task overload, which violates basic, a priori assumptions described by Sternberg (ref. 7), (2) a limited sample size, or (3) insufficient experimental control in progressing from relatively precise laboratory investigations to dynamic simulations.

An analysis of variance (ref. 27) was performed on the RT data from the core experiment employing the RCAH stability augmentation model (table 1) to help identify unknown sources of variation, and to determine the feasibility of pooling selected RT data across selected experimental treatments. Despite skewness in the raw data from the secondary task, ANOVA's were considered sufficiently robust to circumvent the logical problem of running test statistics on data not meeting the model assumptions. As is appropriate for a repeated measures design, F-tests for main effects and selected interactions were recomputed using the next, higher-order interaction as the error term. Strength-of-association values were also estimated for any statistically significant main effect or interaction; they are reported together with the results of the ANOVA in table 7.

From the strength-of-association values (table 7), it is evident that statistically significant findings only account for about 6% of the variation on the secondary task, compared with 39% on the baseline Sternberg condition. Although low strength-of-association values are not unexpected in complex investigations, they failed to provide support for statistically significant ANOVA findings in accounting for much of the total experimental variation. Additionally, three second-order and two higher-order significant interactions precluded pooling the data across the three latter flight segments. Of the main effects, only flight segments (S) approached a statistical level of significance ($F = 5.63$ (df = 3.3) $p < 0.09$). This was probably a result of difficulties in attending to the secondary task during the descent and the hover and bob-up flight segments, as discussed earlier. The difficulty level of the secondary task across flight segments was seemingly affected differentially by the choice of subjects, as well as by the particular controller being flown. The statistically significant, higher-order interactions are difficult to interpret, and no probable explanations are apparent. Thus, pilot FHQ ratings and questionnaire data were examined to identify the factors that contributed to the complicated behavior of the secondary task in this investigation.

Four CHPR's were elicited for each mission segment following each flight (table 8). An ANOVA summary is given in table 9. Of the three main effects, only

TABLE 7.- RESULTS OF ANOVA (ref. 27) PERFORMED ON THE CORE RCAH PORTION OF THIS INVESTIGATION; STRENGTH OF ASSOCIATION VALUES ARE PROVIDED AS AN INDICATION OF TOTAL ACCOUNTED VARIANCE

Source	df	Sum of squares	Mean squares	F	p	Strength of association
Subjects (S)	1	65712.30	65712.30	0.14	N.S.	-
Controller (C)	2	11048081.71	5524040.86	2.38	N.S.	-
Memory set (M)	2	6736264.78	3368132.39	4.08	N.S.	-
Segment (G)	3	27267811.48	9089270.49	5.63	N.S.	-
SC	2	4634441.05	2317220.53	4.77	<.01	0.0065
CM	4	6600664.76	1650166.19	.98	N.S.	-
SM	2	1649452.36	824726.18	1.70	N.S.	-
SG	3	4840061.00	1613353.67	3.32	<.02	.0067
CG	6	13602626.60	2267104.43	9.32	<.01	.0190
MG	6	2407021.77	401170.29	2.91	N.S.	-
SCM	4	6758336.21	1689584.05	3.48	<.01	.0094
SCG	6	1460181.65	243363.61	.50	N.S.	-
SMG	6	827935.84	137989.31	.28	N.S.	-
CMG	12	11784057.13	982004.76	.79	N.S.	-
SCMG	12	14900174.93	1241678.99	2.55	<.01	.0208
Error	1240	602715670.48	486061.02			-

TABLE 8.- COOPER-HARPER PILOT RATINGS OBTAINED ON THE CORE RCAH PORTION OF THIS INVESTIGATION

Controller configuration	Memory set size	Mission segment			
		Descent	Maneuvering	Hover/bob-up	Straight/level
Conventional	0	5/4/5/6	4.5/6/5/5	5/5/4/4.5	4/4/4/3
	1	6.5/4/5/4	5.5/5/5/5	5.5/5/4/5.5	5/4/5/3
	2	5.5/4/5/7	5.5/5/5/4	5.5/6/4/3	5/5/4/3
	4	6/4/5/7	5.5/5/5/8	5.5/5/4/4	5/4/3/3
Four-axis side stick	0	4/5/6/5	6/5/6/4.5	7/5/4/5	5/4/4/3
	1	6/5/6/5	7/5/5/4	8/5/4/5.5	5.5/4/5/3
	2	5/5.5/6/5	7/5/6/4	8/5.5/5/4.5	5.5/4.5/5/3
	4	5/5.5/6/4.5	7/5/6/4	8/5.5/5/5	5.5/4/5/4
Three-axis side stick and collective	0	5/5/5/7	5/4/5/6	6/5/4/4.5	4.5/5/4/4
	1	5.5/5/5/5	6/5/5/6	6/5/4/6	5/6/5/4
	2	5.5/5/6/5	5.5/5/5/6	6/6/4/6	5/5/4/4
	4	5.5/7/5/5	5.5/5/5/6	6/5/4/6	5/6/4/4

Note: The four numbers given in each cell correspond to the four subjects.

TABLE 9.- SUMMARY OF ANALYSIS OF VARIANCE TABLE FOR THE COOPER-HARPER PILOT RATINGS

Source	SS	df	MS	F	P
Mean	4870.2552	1	4870.2552	590.707	0.000
P/	24.7344	3	8.2448		
CONTROL	5.8932	2	2.9466	1.007	.420
CP/	17.5547	6	2.9258		
MEMSET	3.1719	3	1.0573	1.742	.228
MP/	5.4635	9	.6071		
CM	.2422	6	.0404	.122	.992
CMP/	5.9349	18	.3297		
SEGMENT	31.1927	3	10.3976	3.261	.073
SP/	28.6927	9	3.1881		
CS	4.7526	6	.7921	.622	.711
CSP/	22.9245	18	1.2736		
MS	1.6302	9	.1811	.472	.880
MSP	10.3594	27	.3837		
CMS	4.9870	18	.2771	1.017	.457
CMSP/	14.7109	54	.2724		

the flight segment approached a level of statistical significance (F = 3.261 (df = 3,9) $p < 0.075$). Neither memory set size nor control configuration was significant, and there were no higher-order interactions. The tendency for flight segment to approach significance is interesting, since it was the only dependent variable which also approached statistical significance for the Sternberg task. It is probable that a fair degree of pilot compensation was being required to perform the primary flight-control task, and that the secondary task may have been periodically ignored in favor of providing consistent FHQ ratings. The former finding tends to support the earlier contention that excessive loading on the primary task may have inhibited the classical behavior of the Sternberg function.

Each subject was required to complete a postflight questionnaire after each set of runs with a different controller/stability augmentation level combination, as well as a final debriefing questionnaire following the simulation. The following is based on the responses of the four pilot subjects to the postflight questionnaire:

1. Subjects considered the effects of the secondary task harmful to their primary flight-control task, especially while flying three- and four-axis control configurations. Two subjects noted that this was particularly apparent during the descent segment, with conflicting head-up/head-down visual demands. In fact, several pilots initially neglected to respond to the secondary task altogether upon entering the descent and the hover and bob-up segments. One pilot stated that the need to cross-check the secondary task had a profound effect on degrading primary task performance and increasing overall workload. A second pilot stated that the secondary task, especially with the four-axis controller, had a strikingly negative effect on

his ability to analyze handling qualities in flight, and felt that his overall flight performance was compromised.

2. Pilots reported that the need to perform display mode-select control functions had very little effect on their performance on either the primary or secondary tasks.

3. With the exception of one Ames Research Center project pilot, who considered himself to be overtrained, a tendency was noted for the remaining three subjects to consider themselves undertrained on the primary flight-control systems task. As anticipated, this tendency was greatest for the experimental three- and four-axis controllers. Two of the pilots rated themselves extremely undertrained for flying the three-axis control configuration, despite additional training runs.

4. There was a tendency for subjects to rate their performance slightly worse on the secondary task as the number of letters increased in the memory set.

5. There was a decided tendency for subjects to increase their scan pattern difficulty ratings with the addition of the secondary letter recognition task. Two of the four pilots stated that this difficulty was particularly accentuated during descent — a finding undoubtedly related to the head-up/head-down visual demands during that segment.

6. There was a slight tendency to rate the presentation location of secondary task stimuli on the head-up display as nonoptimal. This tendency was related to the head-up/head-down problem during descent, however, and only one subject suggested relocation from the nine o'clock to the twelve o'clock position. Another subject commented that the HUD is a very busy display and must be used during hover and bob-up; addition of the secondary task in this already busy display makes this segment most difficult. However, he further commented that moving secondary task symbols to another display would be unacceptable for this maneuver.

7. Subjects did not indicate that they had made many false identification errors, and for the most part, their perceptions were correct. On several of the most difficult runs, or when subjects fell behind on the primary task, their perception of the number of false identification errors appeared to be inflated. One pilot reported that he had difficulty triggering the voice response mechanism. A small number of such errors were noted, but were discarded from further analysis.

8. As noted earlier, subjects appeared to experience greatest difficulty responding to the secondary task while flying the descent, and pilot ratings on this question tended to substantiate this observation. Difficulties were also identified during hover and bob-up, followed by the NOE segment. Subjects reported only minor difficulties in attending to the secondary task during the straight-and-level flight segment.

9. The consensus regarding the relative difficulty of executing the four flight segments rated the 6° descent as most difficult and straight-and-level as the easiest. Consensus was mixed regarding the NOE and hover and bob-up mission segments, but appeared to be at least partially a function of the flight-control configuration (see appendix A, table 10). Variations in pilot ranking on this question were also related to the amount of exposure to side-arm flight controllers, with specific configurations affecting segment difficulty differentially.

Postflight questionnaires were also completed by two subjects flying the rate-damping stability augmentation model with conventional flight controls. Subject ratings on these questionnaires were compared with their previous ratings flying the same controller, but employing RCAH stability augmentation. Only minor differences were noted between ratings with one possible exception: on question 9, one pilot rated the hover and bob-up segment as most difficult to fly under RCAH, but easiest under the rate-damping stability augmentation level. The 6° descent was rated most difficult to execute under rate damping.

Subjects indicated that they had received adequate training on the secondary task, and on flying the mission profile, but one pilot noted that more training would have been beneficial on the flight controllers, and two pilots indicated that more exposure was needed to the SCAS levels. For the most part, subjects appeared to remain strongly motivated throughout the simulation. Only one pilot indicated some difficulty with fatigue, and another with boredom in flying the same mission profiles. A slight tendency to consider the secondary task distracting from the primary task was again noted, as in the postflight questionnaire results.

CONCLUSIONS

The results of this investigation failed to support the continued use of the Sternberg task as a secondary measure of pilot workload in exploratory FHQ evaluations in which primary task demands may be excessive. Not only did the Sternberg function fail to materialize from data obtained under dual task conditions, but relevant main effects failed to reach levels of statistical significance. Statistical interactions were also largely uninterpretable. Strength-of-association values indicated that only a relatively small portion of the total variance was being accounted for by statistically significant effects. Some attempt was made to understand the otherwise stable and predictable behavior of the Sternberg metric which became erratic in the presence of the primary flight-control task. Possible explanations include the following:

1. Task loading was high on the primary flight-control task. This finding was supported by CHPR data, pilot responses to the questionnaires, and by direct observation of pilot failures to respond to the secondary task. It is likely that pilot responses to the secondary task were inhibited by insufficient reserve capacity owing to excessive demands of the primary task, at least during portions of the mission segments.

2. Insufficient numbers of observations were collected to ensure reliable measures of central tendency for RT's across all experimental variables — a continuing problem in the FHQ test environment in which a relatively small number of highly trained test pilots are typically used. Further decreasing the ISI on the secondary task, or expanding the length of mission segments to obtain a sufficient number of responses, would only tend to degrade performance on the primary task and would likely be unrealistic.

3. Responses to the postflight questionnaire provided some indication of conflict or overload occurring in the visual input channel when the secondary task was presented concurrently with the primary task. Vidulich and Wickens (ref. 16) reported that the Sternberg task was performed most poorly under S-C-R incompatible

REFERENCES

1. Mudd, S. A.: The Treatment of Handling-Qualities Rating Data. Human Factors, vol. 11, no. 4, 1969, pp. 321-330.

2. McDonnell, J. D.: An Application of Measurement Methods to Improve the Quantitative Nature of Pilot Rating Scales. IEEE Transactions on Man-Machine Systems, vol. MMS-10, no. 3, Sept. 1969.

3. Donders, F. G.: On the Speed of Mental Processes. Acta Psychologica 30, Attention and Perception II, W. G. Koster, ed., North Holland Publishing Company, Amsterdam, 1969, pp. 412-4341.

4. Koster, F. G.: Attention and Performance II. North-Holland Publishing Co., Amsterdam, 1969.

5. Sternberg, Saul: High-Speed Scanning in Human Memory. Science, vol. 153, Aug. 1966, pp. 652-654.

6. Sternberg, Saul: Memory Scanning: New Findings and Current Controversies. Quarterly Journal of Experimental Psychology, vol. 27, 1975, pp. 1-31.

7. Sternberg, Saul: The Discovery of Processing Stages: Extension of Donder's Method. Acta Psychologica 30, Attention and Performance II, W. G. Koster, ed., North-Holland Publishing Company, Amsterdam, 1969, pp. 276-315.

8. Wickens, C. D.: The Structure of Attentional Resources. Acta Psychologica 30, Attention and Performance VIII, Raymond S. Nickerson, ed., Lawrence Erlbaum Associates, Publishers, Hillsdale, N.J., 1978, pp. 239-257.

9. Knowles, W. B.: Operator Loading Tasks. Human Factors, vol. 5, 1968, pp. 155-161.

10. Knowles, W. G.; and Rose, D. J.: Manned Lunar Landing Simulation. Proceedings of the IEEE National Winter Convention on Military Electronics, Los Angeles, Calif., Jan. 1963.

11. Wickens, C. D.: Processing Resources in Attention and Workload. TR-EPL-81-3/ONR-81-3, Office of Naval Research, Arlington, Va., July 1981.

12. Micalizzi, J.; and Wickens, C. D.: The Application of Additive Factor Methodology to Workload Assessment in a Dynamic System Monitoring Task. TR-EPL-80-2/ONR-80-2, Office of Naval Research, Arlington, Va., Dec. 1980.

13. Wickens, C. D.; Derrick, W.; Berringer, D.; and Micalizzi, J.: The Structure of Processing Resources: Implications for Task Configuration and Workload. Proceedings of the Human Factors Society, 24th Annual Meeting, Los Angeles, Calif., Oct. 1980.

14. Wickens, C. D.; Sandry, D.; and Micalizzi, J.: A Validation of the Spatial Variant of the Sternberg Memory Search Task: Search Rate, Response Hand, and Task Interference. TR-EPL-81-2/ONR-81-2, Office of Naval Research, Arlington, Va., Mar. 1981.

visual-manual conditions, and best under compatible auditory-speech conditions for both single and dual task conditions. Experimental methods employing an auditory-speech Sternberg task may be more compatible with a heavily demanding visual-manual primary flight-control task.

4. One final factor that merits consideration is the response state used by the pilot subjects during this investigation. Damos (ref. 30) and others have pointed out that strategy is a major determinant of dual-task performance. Chiles (ref. 31) stated that the priority an operator assigns to a task is an important factor in determining the level of performance maintained on that task as other duties are added. Despite careful procedural and experimental controls exercised within the constraints of this investigation, differences in subject strategies and priorities were undoubtedly operative during this study and accounted for at least some of the variability in RT data. It is important to remember that the test pilot's primary job is the FHQ evaluation; it is difficult to modify a learned set of priorities through a limited sequence of instructions and practice trials. The Sternberg task may actually have been relegated to a tertiary position by test pilots as they concentrated on formulating responses to the CHPR scale. It would b interesting and informative to investigate the sequence of events leading to formulation of CHPR's in context of a secondary task.

It is both relevant and informative that CHPR data also failed to provide reliable statistical discrimination between the core (RCAH) treatment conditions in thi investigation. Without quantitative performance or rating data, the FHQ researcher is forced to rely primarily on the pilots' subjective comments and observations regarding flight-control system evaluations. Regardless of the fact that test pilots are trained observers, it is evident that system evaluations can be affected by the number of observers, their background and experience, and other, hard-to-control, intervening factors. The negative findings obtained from this investigation do not obviate the need for continuing research to identify and develop sensitive workload metrics for FHQ investigations, but rather highlight the complex problems and difficulties surrounding this type of work. It is apparent from the current investigation that secondary tasks superimposed upon demanding primary flight-control tasks are likely to yield inconclusive results. Further research efforts are likely to be more productive and acceptable to the FHQ community at large, if an embedded task structure can be defined within the demands of the primary flight-control task. In addition to avoiding the use of additive or more demanding task structures, potential workload metrics should be as unobtrusive as possible from the pilot's viewpoint.

ORIGINAL PAGE IS
OF POOR QUALITY

15. Wickens, C. D.; and Derrick, W.: The Processing Demands of Higher Order Manual Control: Application of Additive Factors Methodology. TR-EPL-81-1/ONR-81-1, Office of Naval Research, Arlington, Va., Mar. 1981.

16. Vidulich, M.; and Wickens, C. D.: Time-Sharing Manual Control and Memory Search: The Joint Effects of Input and Output Modality Competition, Priorities, and Control Order. TR-EPL-81-4/ONR-81-4, Office of Naval Research, Arlington, Va., July 1981.

17. Wickens, C. D.; Sandry, D.; and Vidulich, M.: Compatibility and Resource Competition between Modalities of Input, Central Processing, and Output. Human Factors, vol. 25, no. 2, 1983, pp. 227-248.

18. Crawford, B. M.; Hoffman, M. S.; and Pearson, W. H.: Multipurpose Digital Switching and Flight Control Workload. TR-78-43, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, Dec. 1978.

19. Corrick, G. E.: Missile Launch Envelope Study. TR-80-136, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, June 1981.

20. Johnson, Richard M.: Target Information Processing: The Effects on Reaction Time of Terrain, Downlook Angle, and Response Processing Level. Technical Report 480, Research Institute for the Behavioral and Social Sciences, Army Research Institute, Alexandria, Va., Oct. 1980.

21. Schiflett, S. G.; Linton, P. M.; and Spicuzza, R. J.: Evaluation of a Pilot Workload Assessment Device to Test Alternate Display Formats and Control Handling Qualities. NATO AGARD Conference Proceedings No. 312, Stuttgart, Germany, May 1981.

22. Aiken, E. W.; Blanken, C. L.; and Hemingway, J. C.: A Manned Simulator Investigation of the Effects of an Integrated Isometric Controller on Pilot Workload for Helicopter Nap-of-the-Earth Flight. Presented at the 17th Annual Conference on Manual Control, 15 Oct. 1981, JPL Publication 81-95, pp. 237-239.

23. Schiflett, S. G.: Evaluation of a Pilot Workload Assessment Device to Test Alternate Display Formats and Control Handling Qualities. Report No. SY-33R-80, Naval Air Test Center, NATC, Patuxent River, Md., July 1980.

24. Chen, R. T. N.: A Simplified Rotor System Mathematical Model for Piloted Flight Dynamics Simulation. NASA TM-78575, 1979.

25. Aiken, E. W.: Simulator Investigations of Various Side-Stick Controller/Stability and Control Augmentation Systems for Helicopter Terrain Flight. AIAA Paper 82-1522, San Diego, Calif., 1982.

26. Aiken, E. W.; and Merrill, R. K.: Results of a Simulator Investigation of Control System and Display Variations for an Attack Helicopter Mission. Presented at the 36th Annual National Forum of the American Helicopter Society, Washington, D.C., May 1980.

27. Dixon, W. J.; et al.: Biomedical Computer Programs P-Series. U. of California Press, Los Angeles, Calif., 1981.

28. Hays, W. L.: Statistics for Psychologists. Holt, Rinehart, and Winston, Inc., New York, 1964.

29. Winer, B. J.: Statistical Principles in Experimental Design. McGraw-Hill, Inc., New York, N.Y., 1971.

30. Damos, D. L.: Development and Transfer of Timesharing Skills. TR-ARL-77-11/AFOSR-77-10, Aviation Research Laboratory, U. of Illinois, Urbana, Ill., 1977.

31. Chiles, W. D.: Objective Methods for Developing Indices of Pilot Workload. FAA-AM-77-15, FAA Civil Aeromedical Institute, Oklahoma City, Okla., July 1977.

END

DATE

FILMED

FEB 2 1985